# Structural Determinants of DNA-binding Specificity for Hox Proteins

Peng Liu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

# Abstract


Structural Determinants of DNA-binding Specificity for Hox Proteins

Peng Liu

Hox proteins are a group of homeodomain-containing transcription factors that define the body plan in both vertebrates and invertebrates. Mutations in Hox proteins lead to limb malformations or cancer in humans. Despite having homeodomains with similar sequences and structures, the eight Hox proteins in *Drosophila* exhibit a variety of DNA-binding specificities when they are in complex with their cofactor Extradenticle (Exd), raising the question of how such diverse specificity is generated.

We have identified DNA minor groove shape as a structural determinant for Hox specificity. Using Monte Carlo simulations, we predicted the minor groove widths for Hox-binding sites obtained from a high-throughput experiment - Systematic Evolution of Ligands by Exponential Enrichment with massive parallel sequencing (SELEX-seq). We found that DNA sites selected by anterior Hox proteins have two narrow regions in the minor groove where Hox-Exd binds. In contrast, DNA sites favored by posterior Hox proteins have only one narrow region. Moreover, clustering of Hox proteins based on their preference of DNA minor groove shape reproduced the ordering of Hox genes along the chromosome, suggesting a striking relationship between body axis morphogenesis and nuances in DNA shape.

Intrigued by the question of how DNA shape is recognized, we studied the interactions between an anterior Hox protein, Sex combs reduced (Scr), and its

preferential DNA sites identified from SELEX-seq. Through structure-based homology modeling, we found that two Arg residues on the N-terminal arm of Scr specifically recognize the two narrow regions in the minor groove of Scr-favored sites, regardless of their nucleotide identities. Our work leads to a new understanding of the structural basis of specific DNA-binding for *Drosophila* Hox proteins, linking preference of DNA-binding sites to DNA minor groove shape.

Our studies on Hox-cofactor-DNA structures revealed highly conserved features of protein-DNA recognition, e.g. Hox's Asn51 forms hydrogen bonds to an adenine, which are essential for Hox-DNA binding. In order to automatically identify this type of important interactions, we developed a computational module based on the functional annotation server MarkUs. This module displays a variety of protein-DNA interactions inside query structure and illustrates their degrees of conservation by comparing query structure with its structural homologs. This functional annotation module provides an effective way to analyze protein-DNA recognition and to identify essential interactions.

In this dissertation, Chapter 1 introduces the field of protein-DNA specific recognition from the perspectives of three-dimensional structures, high-throughput experiments, and computational modeling approaches. Chapter 2 introduces the biological background of Hox proteins, focusing on their biological functions, three-dimensional structures, and previous studies on their DNA-binding specificity. Chapter 3 presents the investigation of DNA-binding specificity for Hox-Exd complexes. The role of DNA minor groove width as a structural determinant is demonstrated through Monte Carlo simulations. Chapter 4 describes the homology modeling method for studying DNA minor groove recognition for Scr. The recognition mode of Scr-favored SELEX-

seq sequences is inferred through protein-DNA docking and interface optimization. Chapter 5 elucidates the functional annotation module for protein-DNA structures. The functions and features of this module are demonstrated through a case study on a Scr-Exd-DNA structure. Chapter 6 summarizes my research projects described in this dissertation and proposes future directions for studying specific protein-DNA recognition.

# Table of Contents

# List of Figures and Tables

## Figures

## Tables

# Acknowledgements

First and foremost, I would like to thank my advisor, Prof. Barry Honig, for his support, guidance, and encouragement over the years during my Ph.D. studies. Barry's passion for science, insight on research direction, and way of critical thinking all have set a wonderful example for me, as someone who wants to pursue future career in science. His lab has been an exciting place to develop myself as a scientific researcher. I am particularly grateful for having extraordinary freedom in doing research and the nice working environment.

I am indebted to Prof. Richard Mann, who has co-advised me on the studies of Hox specificity and protein-DNA recognition. Richard has imparted me with a wealthy knowledge of Hox proteins and offered me guidance on the applications of computational methods to experimental biology. His thoughtful, diligent and efficient researching style has provided an excellent role model for me.

I would like to express my gratitude to Prof. Remo Rohs, who introduced me to the field of DNA structures. I have learned a great deal from Remo: not only the ways of doing science, but also his communication skills and team spirit. His love for his family and optimism in life has also been inspiring to me.

I want to thank Prof. Richard Friesner who has served on my dissertation committee during my Ph.D. studies. I also want to thank Profs. Harmen Bussemaker, Tom Tullius, Arthur Palmer, and Larry Shapiro, for their valuable advice and participation in my academic development.

Much of the work described in this dissertation is from the fruitful intra- and inter-lab collaborations. I would like to thank my labmates – Trevor Siggers, Sean West, Markus Fischer, and Nacho Garzon – as well as my colleagues – Matt Slattery, Todd Riley, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Robert Abel, Eric Bishop, and Steve Parker. It has been greatly enjoyable in collaborating with them.

I gratefully acknowledge all the members of the Honig lab. In particular, I would like to thank Katie Rosa, who has offered tremendous help on my Ph.D. studies and my personal life in a foreign country half a globe away from my motherland. It is very nice to have good friends in the lab, especially Klara Felsovalyi, Cliff Zhang, Sean West, and Andy Kuziemko.

I am deeply grateful to my family and friends. My maternal grandparents raised me and taught me how to be a good man. My parents' unwavering support helped me all the way from elementary school to graduate school. I want to thank my wife, Jiaoyang, who has supported and encouraged me through all the years during my Ph.D. studies. I would like to express my gratitude to Zhenhai Zhang, Zhan Yao, and Xiaofei Liu for their enduring friendships.

# Dedication

To my maternal grandparents, my parents, and my wife, for their love, encouragement and support.

# Chapter 1. Background: Protein-DNA Recognition

## 1.1 Introduction

Protein-DNA recognition plays a central role in all aspects of genetic activity in biological system, such as transcription, replication, storage, repair, and recombination. The ability of a protein to bind preferentially to a particular DNA sequence is fundamental to normal cellular activities, morphological development, and response to changing environment (Dynlacht, 1997; Vaquerizas et al., 2009). Therefore, in order to understand the basic functioning principle in biological system, it is essential for us to study the specificity in protein-DNA interaction.

Our knowledge on protein-DNA structures have been increased tremendously since the determination of the first three DNA-binding proteins: λ repressor (Pabo and Lewis, 1982), λ cro (Ohlendorf et al., 1982), and CAP (McKay and Steitz, 1981). Nowadays, there are more than 1,500 structures of protein-DNA complexes have been solved and grouped into over 70 SCOP superfamilies (Rohs et al., 2010). Based on the overall secondary structures of DNA-binding domains, they are categorized as: mainly α, mainly β, mixed α/β, and loops (Garvie and Wolberger, 2001; Rohs et al., 2010). These recognition domains and representative structures will be introduced in Section 1.2.

The enormous protein-DNA structures have increased our knowledge on the origins of specificity in protein-DNA recognition. Traditionally, the specific interaction was interpreted from two aspects: direct readout (Seeman et al., 1976) and indirect readout (Otwinowski et al., 1988). “Direct readout” is defined as the interaction where

DNA sites are recognized by proteins through hydrogen bonds and non-polar interactions between amino acids and bases. “Indirect readout”, on the other hand, refers to interactions when DNA sequences are recognized by proteins through deformed DNA structures. Despite its simplicity, the term “indirect readout” appears to be a loose definition since it encompasses all types of interactions that are not direct. In order to describe the protein-DNA recognition more informatively, two new terms -- base readout and shape readout, were proposed recently to replace the traditional readout mechanisms (Rohs et al., 2010). The detailed concepts and representative structures for these two new terms will be introduced in Section 1.3.

Deep understanding of the underlying principles for specific protein-DNA interaction requires assorted experimental and computational studies. Advance in experimental methods, such as chromatin immunoprecipitation followed by microarray (ChIP-chip) (Ren et al., 2000) or by sequencing (ChIP-seq) (Johnson et al., 2007; Park, 2009), has dramatically increased the repository of DNA-binding sites. For example, ChIP-chip was employed to identify the tissue-specific binding sites for Hox protein Ultrabithorax as well as a Hox cofactor, Homothorax (Slattery et al., 2011a). Moreover, both ChIP-chip and ChIP-seq enable us to determine the DNA-binding sites for different cell types, at different developmental stages, or in different environmental conditions, and therefore provide us plenty of resource to build cellular framework. However, the resolution of experimental measurements is limited to 100 base-pairs, which is not sufficient to identify the exact binding site precisely. Recent technological development on high-throughput methods, such as systemic evolution of ligands by exponential enrichment with massively parallel sequencing (SELEX-seq) (Slattery et al., 2011b),

protein binding microarray (PBM) (Berger and Bulyk, 2009), and bacteria one-hybrid (B1H) (Meng et al., 2005), not only increased the resolution of DNA-binding sites, but also determined the DNA-binding specificity for proteins. These new techniques, together with other high-throughput methods, will be the topic for Section 1.4.

Computational approaches, ranging from bioinformatics analysis to all-atom structural modeling, became as important as experimental approaches towards uncovering the origins for specificity in protein-DNA recognition (Bulyk, 2003; Rohs et al., 2009a). Bioinformatics analysis, an indispensable tool for high-throughput methods, is used to compute both the nucleotide identity and associated affinity for DNA-binding sites. Valuable insights have been obtained and advanced algorithms have been developed to predict specificity on a genome-wide scale. All-atom modeling, on the other hand, decodes specificity from the three-dimensional structural level. It unravels the biophysical basis of specificity and applies them to decipher the recognition code for protein homologs and engineer DNA-binding domains to derive proteins with novel specificity. Furthermore, all-atom modeling has been combined with bioinformatics analysis to refine genome-wide transcription factor binding sites and assign cis-regulatory element to structural families. Section 1.5 will describe the details of bioinformatics analysis and all-atom modeling approaches.

## 1.2 Structural families of DNA-binding domains

Proteins recognize DNA through all three types of secondary structure elements: α-helix, β-sheet, and loop. In this section, I will introduce protein families which use each

of these three elements to contact DNA, elucidate their structural features, and describe representative structures.

### 1.2.1 **α-helix**

The α-helix is the mostly used secondary structural element to recognize DNA sites, mainly through binding to DNA major grooves. When α-helix is inserted into DNA major groove in parallel to the flanking DNA backbones, it will create ideal geometry for protein side-chains to establish hydrogen bonds and non-polar interactions with bases. This type of recognition has been observed in many protein families: such as $Cys_2His_2$ ($C_2H_2$) zinc finger (Wolfe et al., 2000), homeodomain (Gehring et al., 1994a), basic region-leucine zipper (bZIP) (Ellenberger, 1994), etc. Another role for α-helix in protein-DNA recognition is to insert into DNA minor groove, represented by Lac repressor (Lewis et al., 1996).

$C_2H_2$ zinc finger composes the largest transcription factor family in human genome (Vaquerizas et al., 2009). These DNA-binding proteins usually contain multiple copies of a compact, about 30 amino acids domain. Each domain has a relatively short α-helix for inserting into DNA major groove, two anti-parallel strands of β sheet, and a $Zn^{2+}$ ion coordinating two Cys and two His residues (Figure 1-1 A). Multi-domains are connected by short linker regions so that successive fingers bind DNA major groove in a manner of center-to-center spacing. Canonical structures for $C_2H_2$ zinc finger family are Zif268 (Pavletich and Pabo, 1991) and Gli (Pavletich and Pabo, 1993) with three and five DNA-binding domains, respectively.

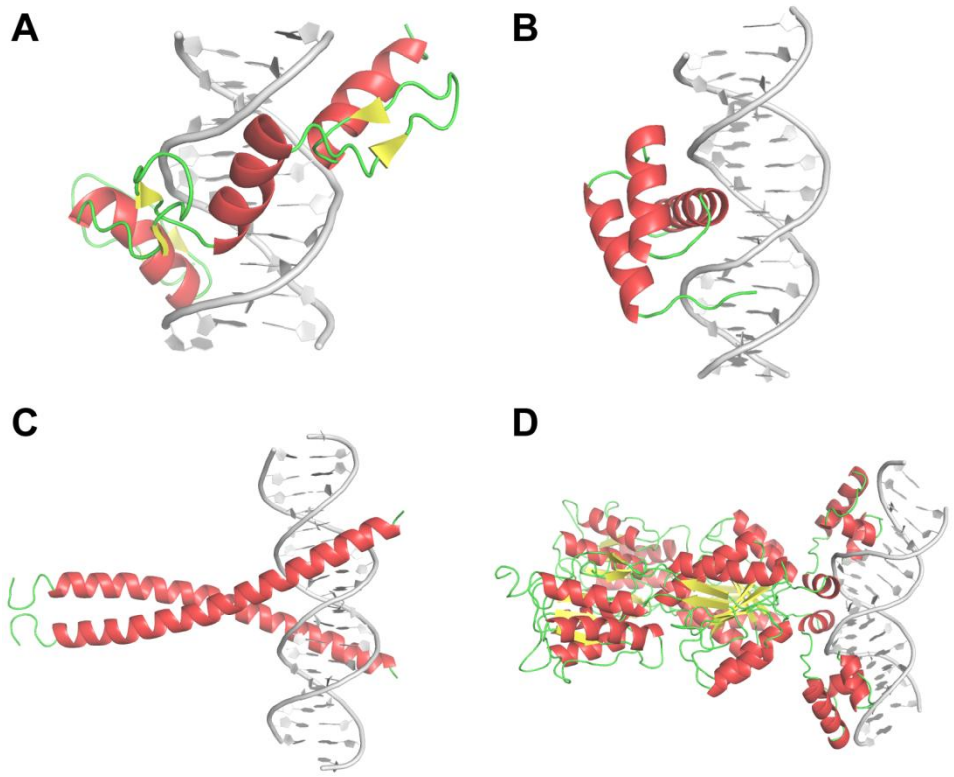


**Figure 1-1. Proteins use α-helices to contact major and minor grooves of DNA.**

(A) $C_2H_2$ zinc finger protein (PDB ID: 1aay); (B) homeodomain (PDB ID: 9ant); (C) bZIP (PDB ID: 1dgc); (D) purine repressor (PDB ID: 1pnr).

Homeodomain is the second largest transcription factor family in human genome (Vaquerizas et al., 2009). Its DNA-binding motif is a conserved bi-helices defined as helix-turn-helix (HTH). In HTH, the second helix inserts into DNA major groove to make base and sugar-phosphate backbone contacts (Figure 1-1 B). The first helix does not embed in the major groove, but make additional DNA contacts in some cases. The two helices are connected by a tight bend and positioned in a relatively fixed angle. The N-terminal loop region of homeodomain, termed N-terminal arm, can function to contact DNA minor groove. It is very flexible when homeodomain binds to DNA alone (Gehring

et al., 1994a), but becomes stabilized and inserts into DNA minor groove when homeodomain binds to DNA together with cofactors (Mann et al., 2009).

The bZIP is also a large family of eukaryotic DNA-binding proteins (Vaquerizas et al., 2009). Members of this family usually bind to DNA as homodimers or heterodimers, which expand the repository of DNA sequences to be recognized. Protein from bZIP family consists of a long helix with about 60 amino acids (Figure 1-1 C). The N-terminal half is the dimerization region and the C-terminal half is inserted into DNA major groove. A notable feature of bZIP proteins is that the C-terminal half is disordered in absence of DNA. The basic region-helix-loop-helix (bHLH) family shares a similar mode of DNA recognition with bZIP family (Nair and Burley, 2003). The major difference between these two families lies in the N-terminal region, where bHLH consists of two helices connected by a loop.

The LacI family consists of proteins with α-helix bind to DNA minor groove. As represented by the purine repressor dimer (Schumacher et al., 1994), each monomer uses a two-turn "hinge" helix to contact minor groove (Figure 1-1 D). To accommodate the two helices, minor groove deforms from the standard B-DNA conformation and becomes 45 ° kinked and underwound. The intercalation of protein side-chains can facilitate the kink as observed in the structure of purine repressor dimer.

### 1.2.2 **β-sheet**

Despite not as widely used as α-helix, β-sheet has been revealed as another type of secondary structural element to contact DNA in both major and minor grooves. The MetJ repressor (Somers and Phillips, 1992), for example, binds to DNA major groove

with an anti-parallel β-sheet, where each monomer contributes a single strand (Figure 1-2 A). The β-sheet is parallel to the flanking sugar-phosphate backbone, allowing contacts formed between protein side-chains and DNA bases.

In contrast, insertion of β-sheet into minor groove requires deformation of DNA structures. Proteins from the TATA-binding protein (TBP) family use a ten-strand β-sheet bind to DNA minor groove (Figure 1-2 B) (Kim et al., 1993a; Kim et al., 1993b). Similar to the Lac repressor site, the DNA undergoes bending and unwinding to accommodate β-sheet and intercalation is also used to promote deformation of DNA minor groove.


**Figure 1-2. Proteins use β-sheets and loops to contact DNA.**

(A) Met repressor (PDB ID: 1cma); (B) TBP (PDB ID: 1ytb); (C) runt domain (PDB ID: 1hjc).

### 1.2.3 **Loop**

In addition to α-helix and β-sheet, loop is also used as a structural element to recognize DNA site. Members from the SCOP family (Andreeva et al., 2008) p53-like transcription factors, such as Rel/Dorsal (Ghosh et al., 1995), p53 (Kitayner et al., 2010), and runt domain (Tahirov et al., 2001), all recognize DNA through loops (Figure 1-2 C). They share a β-sheet immunoglobulin-like domain, but are different in the binding orientation when their loops contact DNA. Structure elements outside the

immunoglobulin-like domain diverge dramatically among family members. A striking feature is that the immunoglobulin-like domain not only recognizes DNA through its loop, but also mediates homo- and hetero-dimer interactions as observed from Rel/Dorsal domain.

## 1.3 Structural determinants for specificity in protein-DNA recognition

Recently, the terms “base readout” and “shape readout” has been coined to describe the origins for specificity in protein-DNA recognition (Rohs et al., 2010). These two new terms were proposed to replace the traditional definitions of “direct readout” and “indirect readout”, mainly because the vague meaning of “indirect readout”, which encompasses all types of interactions that are not direct. In this section, the structural mechanisms and representative structures for base readout and shape readout will be introduced. A combined use of both readout mechanisms in higher-order protein-DNA complexes, such as nucleosome and integration host factor (IHF) will be described as well. One thing to note is that although there are studies have been carried out on the transient protein-DNA recognition (Blanco and Montoya, 2011; Fuxreiter et al., 2011), here we mainly focusing on introducing the “stable” protein-DNA interactions.

### 1.3.1 **Base readout**

Base readout is defined as proteins recognize DNA bases by their chemical signatures in either major or minor groove. This type of readout is carried out mainly through direct hydrogen bonds, water-mediated hydrogen bonds, or hydrophobic contacts, which convey the highest degree of specificity.

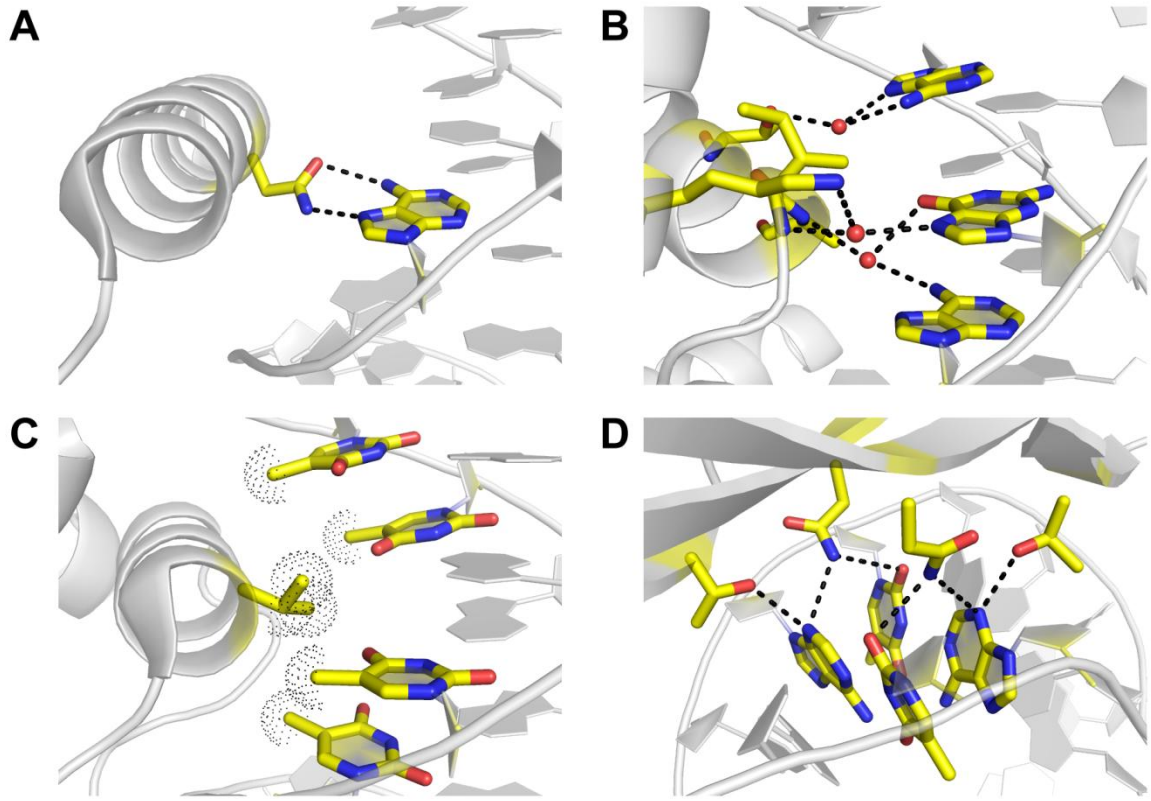


**Figure 1-3. Base readout mediated by hydrogen bonds or hydrophobic interactions.**

(A) Homeodomain (PDB ID: 2r5z); (B) Trp repressor (PDB ID: 1tro); (C) P22 c2 repressor (PDB ID: 2r1j); (D) TBP (PDB ID: 1ytb). Hydrogen bonds are represented by dashed lines. Water molecules are denoted as red sphere. Hydrophobic interactions are show in dotted spheres.

Protein-DNA recognition by direct hydrogen bonds in major groove has been observed in many protein families, including $C_2H_2$ zinc finger (Wolfe et al., 2000), homeodomain (Gehring et al., 1994a), and bZIP (Ellenberger, 1994). The specificity is achieved mainly through bidentate hydrogen bonds between pairs of Arg-guanine, Asn-adenine, or Gln-adenine (Figure 1-3 A). Guanine, for example, is the only base that has two hydrogen bond acceptors in the major groove. This unique chemical feature enables guanine to be specifically recognized by Arg, which have side-chain with multiple

hydrogen bond donors. Arg-1 in $C_2H_2$ zinc finger and Asn51 in Hox homeodomain both use bidentate hydrogen bonds to recognize specific bases in their DNA-binding sites.

In DNA major groove, water-mediated hydrogen bond is another type of base readout, because water molecules often reflect the positions of hydrogen bond donors and acceptors at base edges. This type of readout has been observed in Trp repressor (Otwinowski et al., 1988), retinoid X repressor (RXR)-retinoid acid repressor (RAR) (Rastinejad et al., 2000), and many enzymes (Tainer and Cunningham, 1993). In Trp repressor, three ordered water molecules mediate hydrogen bonds between protein residues and DNA bases, appearing to be important for Trp repressor's specificity (Figure 1-3 B). Similarly, at the RXR-RAR-DNA interface, several Arg/Lys-base interactions involved water-mediated hydrogen bonds.

Hydrophobic contact is a third type of base readout mechanism in major groove. It is mainly used to differentiate thymine from cytosine. Lambdoid bacteriophage P22 c2 employs a Val residue to specifically recognize four consecutive thymine methyl groups (Figure 1-3 C) (Watkins et al., 2008). Similar specific hydrophobic interactions have also been observed in bacteriophage 434 repressor (Aggarwal et al., 1988) and 434 Cro protein (Wolberger et al., 1988).

All three categories of base readout mechanisms have also been employed in the minor groove, although to a much less extent than in the major groove. The main reason is due to limited chemical properties to distinguish each base. For instance, the pattern of donors and acceptors is insufficient to differentiate AT from TA or GC from CG base pairs. Nevertheless, direct hydrogen bonds in minor groove have been found to play a

role in TBP (Figure 1-3 D) and high-mobility group (HMG) proteins (Huth et al., 1997). In addition, hydrophobic contacts have also been employed by TBP to recognize the completely dehydrated TATA box (Kim et al., 1993a; Kim et al., 1993b).

### 1.3.2 **Shape readout**

Shape readout is carried out through two types of mechanisms: local shape readout and global shape readout. Local shape readout refers to recognize DNA structures that are deviated from ideal B-DNA structure, such as narrow minor groove over three to eight base pairs, DNA kinks, and intercalations. In contrast, global shape readout describes the scenario where the entire DNA-binding site is bent or in A/Z-form helix instead of classic B-DNA conformation.

Recognition of DNA narrow minor groove is a novel concept that was discovered several years ago from *Drosophila* Hox protein Sex comb reduced (Scr) (Joshi et al., 2007). Scr's *in vivo* specific site *fkh250* has an extra narrow region in minor groove than Scr's non-specific site *fkh250*$^{con}$. It is this extra narrow region that induces enhanced electrostatic potential, thereby attracting Scr's Arg3 and His-12 to bind (Figure 1-4 A) (See Section 2.3.2 for more details.). A later survey on all available structures revealed that this type of narrow minor groove recognition is a readout mechanism used by many SCOP superfamilies (Rohs et al., 2009b). A variety of proteins employ this type of narrow minor groove recognition to preferentially bind to their specific DNA sites, such as the pathogen transcriptional repressor MogR (Shen et al., 2009), the bacterial nucleoid-associated protein Fis (Stella et al., 2010), the H-NS related protein Ler (Cordeiro et al., 2011), and the Smad4 MH1 domain (Baburajendran et al., 2011). The tumor suppressor protein p53, in particular, recognizes its DNA site through local shape,

where a Hoogsteen base-pair reduces the helix diameter and narrows the minor groove of the flanking regions (Kitayner et al., 2010). Hoogsteen base pairs have been observed in free DNA structures (Nikolova et al., 2011) and could potentially offer new ways for protein to recognize its binding site (Honig and Rohs, 2011).

**Table 1-1. Desolvation energy of ionized arginine and lysine side-chains.**

| Side-chain | Desolvation energy from ε=80 to ε=2 [kcal/mol] | | | |
|---|---|---|---|---|
| | AMBER94 | CHARMM | OPLS | PARSE |
| Lysine | 36.53 | 39.27 | 36.98 | 41.10 |
| Arginine | 34.24 | 34.58 | 30.39 | 35.20 |
| Difference | 2.29 | 4.69 | 6.59 | 5.90 |

This table is adapted from the Supplementary Table 4 in a published paper (Rohs et al., 2009b).

**Table 1-2. Number of hydrogen bonds between DNA minor groove and side-chains of arginine or lysine.**

| | Average | Standard deviation |
|---|---|---|
| Arginine | 0.92 | 0.92 |
| Lysine | 0.61 | 0.63 |

This table is adapted from the Supplementary Table 3 in a published paper (Rohs et al., 2009b).

Among the twenty types of amino acids, arginine is the most frequently used one to recognize enhanced negative electrostatic potential in DNA's narrow minor groove (Rohs et al., 2009b). By contrast, the other positively charged residue, lysine, appears much less. The main reason accounting for this difference is in desolvation energy (Rohs et al., 2009b). Theoretical calculations based on four different force fields showed that lysine requires 2.3~6.6 kcal/mol more energy to desolvate than arginine (Table 1-1). Although earlier study attributed arginine's enrichment to the ability of arginine's guanidinium to make more hydrogen bonds than the amino group of lysine (Luscombe et al., 2001), a statistical survey revealed that on average arginine only make a slightly

higher number of hydrogen bonds than lysine (0.9 vs. 0.6) and the large standard deviation indicates that such difference is insignificant (Table 1-2). Therefore, it is the desolvation energy, not the number of hydrogen bonds, makes arginine more enriched in the recognition of narrow minor groove than lysine.

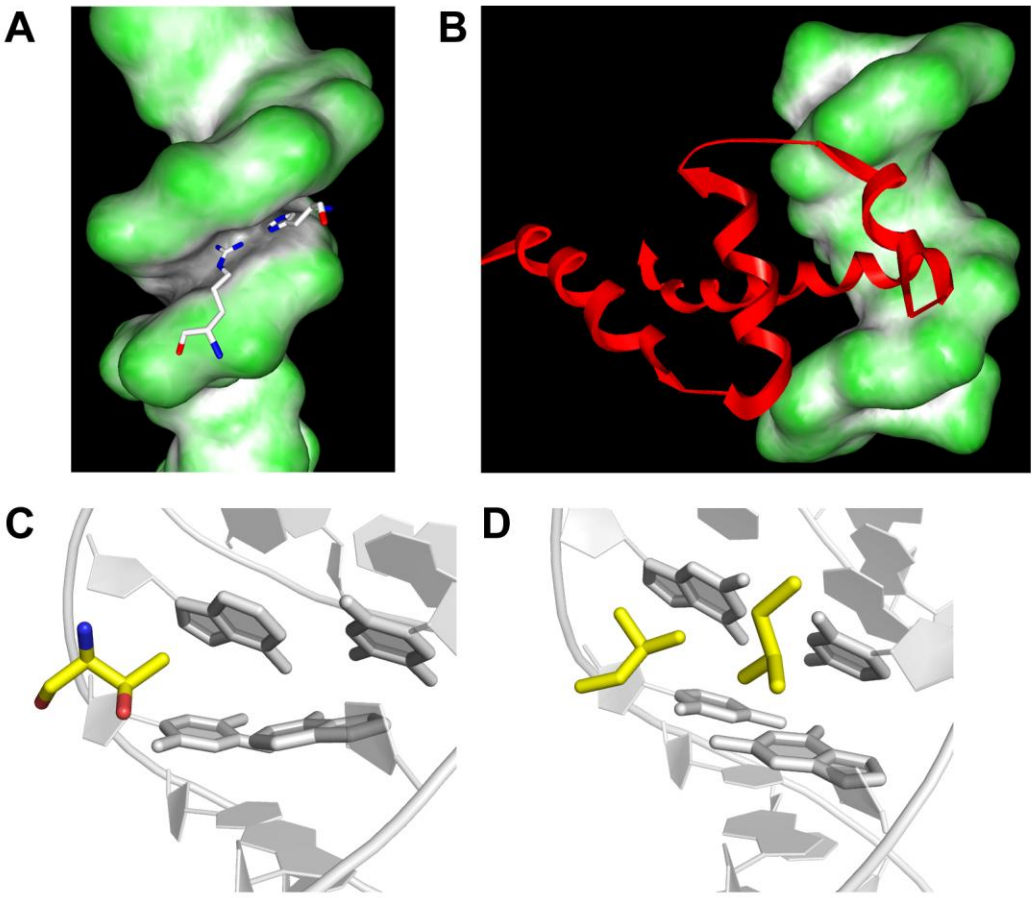


**Figure 1-4. Examples for local shape readout.**

(A) Recognition of narrow minor groove in Scr (PDB ID: 2r5z); (B) Recognition of wide major groove in $\sigma^E$ factor (PDB ID: 2h27); (C) Kinks in γδ resolvase (PDB ID: 1gdt); (D) Intercalation in Lac repressor (PDB ID: 1l1m; figure is made from the first one of the 20 NMR conformations).

Similar to minor groove, the width of DNA major groove can also be specifically recognized by proteins, considering that the shape of minor and major groove correlates

with each other. The group IV $\sigma^E$ factor from *E. coli*, for instance, binds preferentially to 5'-GGAACTT-3', where ApA step is highly conserved. Structural study uncovered that ApA step corresponds to a narrow region in minor groove and therefore increases the shape complementarity in the major groove (Figure 1-4 B) (Lane and Darst, 2006).

Kinks and intercalations are another two categories of local shape readout. Kinks refer to DNA structure where the linearity of helix is abruptly broken due to flexible base pair step, such as TpA. This unique geometry of kink can optimize protein-DNA and protein-protein contacts, therefore contributing to the binding specificity. The binding site of γδ resolvase contains a 5'-TATA-3' motif that has sharp kinks at both TpA steps, allowing resolvase to bind in a favorite geometry (Figure 1-4 C) (Yang and Steitz, 1995). Intercalation often occurs to stabilize kink in DNA by inserting protein side-chains between base pair step. As introduced in the previous section, the Lac repressor has two Leu residues intercalated into kinks, respectively, as a way to recognize two CpG steps (Figure 1-4 D) (Kalodimos et al., 2004).

Global shape readout always involves the recognition of the entirely deformed DNA sites, either bent or in A/Z-form helix. Human papillomavirus HPV-16 and HPV-18 E2 proteins have a strong preference of 5'-AATT-3' in the linker region over 5'-ACGT-3', although none of the base-pairs is in contact with proteins (Hegde, 2002). Structural and Monte Carlo studies revealed that E2-binding sites with 5'-AATT-3' as a linker is intrinsically bent, whereas site with 5'-ACGT-3' is straight (Rohs et al., 2005b). Global shape readout can also be carried out through the recognition of A-DNA or Z-DNA. In A-DNA, sugars are exposed in minor groove and thereby provide large hydrophobic contacting surface, such as in the TBP-TATA boxes binding (Guzikevich-Guerstein and

Shakked, 1996). Z-DNA has zigzag positioning of phosphates along left-handed helices, which can be specifically recognized by RNA adenosine deaminase and tumor-associated DLM-1 protein (Schwartz et al., 2001).

### 1.3.3 **Higher-order complexes with combined readout mechanisms**

Base readout and shape readout provide a reductionist view on the origins of specificity in protein-DNA recognition. However, in most cases, a combination of assorted readout mechanisms is used to achieve specificity. The nucleosome complex in eukaryotes and IHF complex in *E. coli* are two typical examples.

Nucleosome is the key DNA packing unit in eukaryotes, consisting of about ~147 base-pairs DNA wrapped around histone proteins (Andrews and Luger, 2011; Luger et al., 1997). The positioning of nucleosome along genome affects the activity of many transcription factor and other DNA-binding proteins (Segal and Widom, 2009). The DNA sequence alone was indicated to determine the nucleosome occupancy and removal (Wang et al., 2011). A variety of shape readout is found in the histone-DNA interactions. The histone-bound DNA has a periodic, about ten base-pairs on average, narrow minor groove resulted from short A-tracts. Most of these narrow minor regions are presumably recognized by Arg from histones (Rohs et al., 2009b). Moreover, the periodic A-tracts together with kinks of CpA steps enable the bending of DNA, therefore enhancing the fitness of overall DNA curvature to wrap around histones (Olson and Zhurkin, 2011; Tolstorukov et al., 2007).

IHF is an architectural protein in *E. coli*, maintaining DNA super-coiling and compaction (Rice et al., 1996). It employs a combination of bending, kinks, and

intercalations to recognize the preferential DNA sequences. The 35 base pairs IHF-binding site is bent by more than 160 ° to facilitate wrapping around IHF within a very short distance (Figure 1-5 A). Two large kinks occur at the ApA step in IHF-bound DNA, allowing two highly conserved Pro to insert, respectively (Figure 1-5 B).

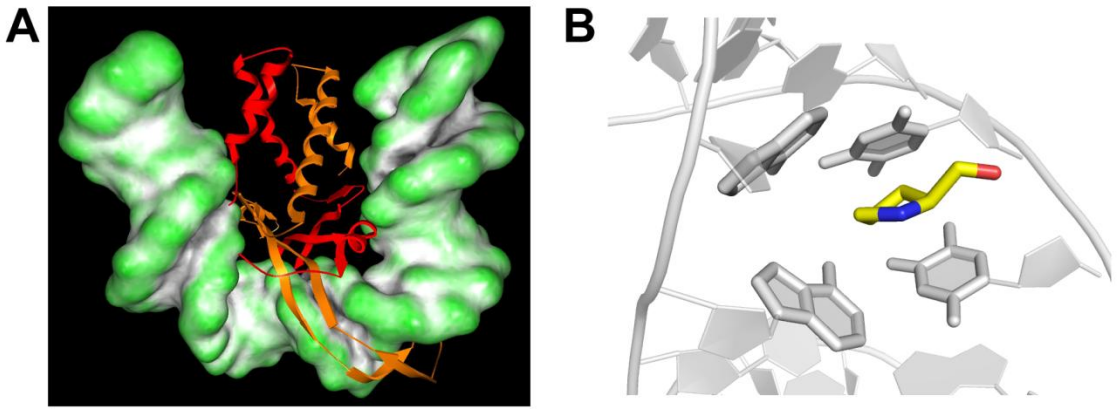


**Figure 1-5. IHF uses a combined readout mechanisms to recognize DNA.**

(A) Global bending of DNA; (B) A large kink recognized by Pro. The PDB ID for this structure is 1ihf.

## 1.4 High-throughput methods for determining specificity in protein-DNA recognition

In the past several years, advances in high-throughput technology have dramatically increased the repositories of *in vitro* and *in vivo* DNA-binding sites for a variety of proteins. In this section, I will introduce three high-throughput methods: Systematic Evolution of Ligands by Exponential Enrichment with massive parallel sequencing (SELEX-seq), bacterial one-hybrid selections, and protein-binding microarray, which all have been applied to study the specificity of Hox proteins. First, I will focus on their experimental protocols. (For applications of these three methods on

Hox proteins, please see Section 2.4.) Then, I will briefly summarize other new high-throughput methods, such as cognate site identifier, mechanically induced trapping of molecular interactions, and surface plasmon resonance.

### 1.4.1 **SELEX-seq**

The SELEX-seq method has been applied to identify preferential DNA sites for eight *Drosophila* Hox proteins in complexes with their cofactors Exd and Hth (simply referred together as Exd). Novel insights were obtained on the latent specificity of Hox proteins evoked by cofactors (Slattery et al., 2011b).

The SELEX-seq method combined Systematic Evolution of Ligands by Exponential Enrichment with massively parallel sequencing (Figure 1-6). Starting from a pool of synthesized DNA oligonucleotides containing a region of 16 random base-pairs, EMSAs are performed and DNA sites bound to Exd-Hox heterodimers are isolated and amplified by PCR. The affinity-based selection step is repeated twice. The initial pool and the enriched pool from each round are sequenced. Relative binding affinities for DNA sites are calculated based on the round-round enrichments. To avoid bias in sequence composition, a Markov model is constructed to predict the relative frequency of all 16-mers in the initial pool and a LOESS regression is employed to combine information from earlier and later rounds of selection.

The SELEX-seq method provides an ideal framework to study specificity in protein-DNA interactions at a large scale. EMSA enables researchers to focus on a heterodimeric complex. The use of Markov model and LOESS regression makes it

possible to compute relative affinities for DNA sites, which have abundance differed by almost two orders of magnitude between two rounds of selection.

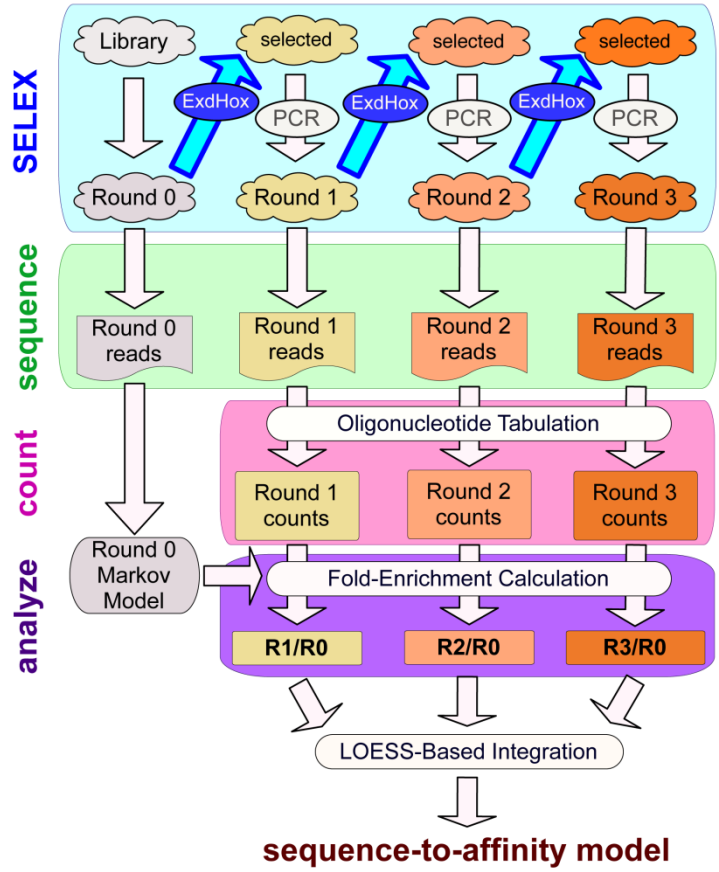


**Figure 1-6. Overview of the SELEX-seq method.**

Multiple rounds of selection for Exd-Hox bindings sites, massively parallel sequencing, and a biophysical model to compute relative binding affinity. Figure is adapted from a published paper (Slattery et al., 2011b).

### 1.4.2 Bacterial one-hybrid (B1H) selections

The B1H selections provide a simple and rapid way to determine DNA-binding specificity of a transcription factor (TF) (Christensen et al., 2011; Meng et al., 2005). The DNA binding domain (DBD) of a given TF is fused to the α-subunit of RNA polymerase

and the random DNA sites is inserted into the upstream of a weak promoter, which drives the expression of two yeast reporter genes, *HIS3* and *URA3*. If DBD binds to the target DNA site, it will recruit RNA polymerase to the promoter and thereby activate transcription of *HIS3* and *URA3*. DNA site recognized by a given DBD is selected by a combination of positive selection in present of the DBD and a negative selection without the DBD. First, reporter vectors with a binding site for the DBD is selected by growing on minimal medium containing 3-amino-triazole (3-AT), a competitive homolog of *HIS3*. Then, reporter vectors harboring DNA site that actives the promoter independent of the DBD are removed by growing on medium containing 5-fluoro-orotic acid (5-FOA), which can induce toxicity by the uracil biosynthesis pathway.

The advantages of B1H method is that the DNA-binding protein does not need to be purified or synthesized *in vitro*. Moreover, there is no restriction on the length of DNA-binding sites to be selected. Given the efficiency of transformation of DNA site into *E. coli*, it is possible to identify binding site as long as 12-base pairs. However, this method is not suitable for every TF, due to differences in codon bias, folding problems, and toxicity. In addition, confounded by the selection for cell growth over many generations, the relative binding affinity is not computed as straightforward as SELEX-seq. Growth rate needs to be simply assumed proportional to the TF occupancy of the DNA site. Nevertheless, B1H selection offers a low-tech and powerful system to determine DNA-binding specificities for TFs.

### 1.4.3 **Protein-binding microarray (PBM)**

PBM was originally developed to characterize *in vitro* binding specificities of TFs about a decade ago (Berger and Bulyk, 2009; Bulyk et al., 2001). It has dramatically

expanded the repository of binding specificities for TFs (De Masi et al., 2011; Gordon et al., 2011). Current microarray contains over 44,000 spots that allow all possible ten base-pairs DNA sites occur more than 30 times. The PBM experiment is generally performed by first adding epitope-tagged TF to the array, then washing the array to remove nonspecific binding and labeling with fluorophore-conjugated antibody. Analysis of fluorescence on the array provides a quantitative readout of the relative amounts of TF bound to each spot and thereby derives TF's DNA-binding specificity.

A striking feature of PBM is that the relative affinities of a TF for tens of thousands of individual DNA sequences are measured simultaneously. In this way, the results can be obtained in less than a day, which is much faster than other *in vitro* methods. Furthermore, this expeditious advantage allows PBM to provide insights into the *in vivo* activities of TFs. DNA-binding sites for a large number of TFs can be measured across various cellular states and environmental perturbations. On the other hand, PBM is restricted by the number of sequences allowed to be represented on a microarray. Besides, the *in vitro* nature of PBM compromises its application for identifying functional sites *in vivo*. Even though, PBM constitutes a significant paradigm in generating comprehensive binding data for TFs.

### 1.4.4 **Other high-throughput methods**

Cognate site identifier (CSI) (Warren et al., 2006) is a similar method to PBM. It determines the specificity of TFs by measuring their DNA-binding sites for as long as ten base pairs. The major difference between CSI and PBM is the way to prepare initial DNA sites on the microarray. CSI first synthesizes single-stranded DNA and then folds them back to make double-stranded DNA sites. It does not need the primer sequence, which is

required for PBM. CSI has been applied to obtain a full DNA-binding profile for an engineered polyamides molecule PA1 and a Hox cofactor Exd. Combined with a data analysis method named Sequence-Specificity Landscape (SSL), CSI was demonstrated to be a powerful tool to understand the role of non-conserved flanking sequences on binding site (Tietjen et al., 2011).

Mechanically induced trapping of molecular interactions (MITOMI) is a high-throughput micro-fluidic platform that can determine the DNA-binding affinity directly (Maerkl and Quake, 2007). The arrangement of TFs and random DNA sites is inversed to PBM. TFs are attached to the surface of the platform through antibodies, and then synthetic DNA sites are flowed into the chamber. Mechanical trapping and washing are carried out to remove molecules that are not bound to TFs. The amount of TFs and bounded DNA at different concentrations are determined by fluorescent signal, and thereby the relative dissociation constant between TFs and their binding sites can be derived. MITOMI was applied to four basic helix-loop-helix TFs, for which binding affinities for 464 different DNA sites were determined. Potential *in vivo* regulatory roles for these TFs were proposed base on the MITOMI measurement.

Surface plasmon resonance (SPR) was originally developed for characterizing protein-ligand and protein-protein interactions. It has also been applied to determine the DNA-binding specificity of proteins (Campbell and Kim, 2007). Like PBM, SPR has DNA sequences attached to the surface and TF added to the solution so that TF can bind to the DNA to reach equilibrium. Then, the array is washed to dissociate TF and its DNA-binding site. Since the angle of light reflected from the surface correlates with the mass of molecules attached on the other side, the on and off binding rate can be

determined and thereby derived the binding affinity. Similar to PBM, SPR can identify multiple binding sites for the same TF simultaneously, thus it can be used to generate the DNA-binding profile for TFs efficiently.

Bind-n-Seq is a high-throughput method similar to SELEX-seq (Zykovich et al., 2009): incubate TF with random oligonucleotides, sequence the DNA-binding sites with massive parallel technology, and identify the binding motifs. Bind-n-Seq has been applied to determine the binding sites and relative affinities for two $C_2H_2$ zinc finger proteins: Zif268 and Aart.

## 1.5 Computational approaches for determining specificity in protein-DNA recognition

### 1.5.1 **Bioinformatics approaches**

A comprehensive analysis of free DNA and protein-DNA structures has found that the recognition of DNA narrow minor groove is a widely used mechanism for many protein superfamilies (Rohs et al., 2009a). When DNA site contain an A-tract, a run of at least three base-pairs of As or Ts except TpA step, its minor groove always becomes narrow. The narrow minor groove induces enhanced electrostatic potential through electrostatic focusing (Honig and Nicholls, 1995), and thereby attracts protein residues, mainly Arg, to bind. In this way, protein reads DNA sequence by recognizing DNA local shape through electrostatic interactions. The nucleosome core complex presents a prominent example for this mechanism. In the 147 base pair DNA fragment, there are 9 out of 14 narrow regions having Arg in contact and a similar repeating pattern was seen

in 35 nucleosome crystal structures available at that time. Moreover, a periodicity of A-tracts was observed from *in vivo* yeast nucleosome sequences.

Bioinformatics methods built on hydroxyl radical cleavage data (Parker et al., 2009) and uranyl photo-cleavage data (Mollegaard et al., 2005) were developed to understand DNA minor groove shape readout. Both types of cleavage data were demonstrated to correlate with DNA minor groove widths. The two bioinformatics methods have been applied to predict DNA shape for genomic DNA sequences and revealed periodic structural signals related to nucleosome positioning (Bishop et al., 2011; Lindemose et al., 2011). The method based on hydroxyl cleavage data was also applied to genomic sequences from multiple species and found that DNA shape is even more conserved than DNA sequence (Parker et al., 2009; Parker and Tullius, 2011).

Many bioinformatics approaches based on sequence alignments have also been employed to characterize nucleosome's specificity (Segal and Widom, 2009). A preference of periodicity about ten base-pairs on ApT/TpA/ApA steps was observed based on the alignments of *in vivo* nucleosome sequences from many model organisms, including yeast, fly, chicken, and human (Segal and Widom, 2009). A computational model, which is based on direct measurements of nucleosome occupancy assembled on purified yeast genomic DNA, predicted the preferences of nucleosome sequence with a high per-base-pair correlation of 0.89 to experimental results. The prediction results resembled many features of nucleosome's *in vivo* binding sites in yeast and worm, suggesting that DNA sequence is the major determinant for nucleosome specificity (Kaplan et al., 2009). In addition to sequence alignments, helical parameters, which describe inter-base pair translations and rotations, were also employed to study

nucleosome positioning. The study revealed that a superhelical pitch leads to DNA deformation and thus facilitates wrapping around histones (Olson and Zhurkin, 2011; Tolstorukov et al., 2007).

### 1.5.2 **All-atom computational simulations for predicting DNA structures**

Structural knowledge of sequence-dependent DNA conformation is limited because of insufficient X-ray and NMR structures for free DNA (Rohs et al., 2009a). Solving a free DNA structures with X-ray crystallography is often more difficult due to DNA's high flexibility. Lack of NOE constraints confounds NMR method to solve free DNA structures. These disadvantages of experimental methods make all-atom computational simulations an alternative approach to study sequence-dependent DNA conformations.

Given the different algorithms on conformational sampling, all-atom simulations can be divided into two categories: molecular dynamics (MD) simulations where molecular trajectories are calculated by Newton's equations of motion, and Monte Carlo (MC) simulations which rely on repeated random sampling. Both methods depend on force fields that origin from quantum-mechanical calculations, to compute the conformational energies of molecules. For simulations on free-DNA structures, AMBER is the only extensively tested force field (Cheatham, 2004). It has been applied to predict the structures of Drew-Dickerson dodecamer, DNA bending, and A/B-DNA transition (Cheatham and Kollman, 2000). All these applications were shown to agree qualitatively with experimental results. However, systematic under-twisting was observed in MD simulations based on AMBER94 due to the α/γ-flips artifacts. Improvement on AMBER force field has been carried out over the years. AMBER99 were reported to generate a

less under-twisting in MD simulations (Fujii et al., 2007). The Barcelona modification, which further refined torsional parameters, produced more stable simulations (Perez et al., 2007) and generated models with intrinsic bending for sequences containing an $A_4T_4$ tract (Lankas et al., 2010). Another force field CHARMM has also been applied to study free DNA structures. Although distinct sequence-dependent DNA conformations were observed on the activator and repressor motifs for *Drosophila* transcription factor Dorsal (Mrinal et al., 2011), long simulations need to be carried out to test its performance (Orozco et al., 2008).

MC simulations as an alternative approach to MD have also been applied to investigate free DNA structures. A new MC method was developed a few years ago to simulate free DNA conformation by using internal and collective variables derived from the chemical topology of DNA, explicit ions, and a slightly modified version of AMBER94 with a screened Coulomb potential (Sklenar et al., 2006). This algorithm is effective in sampling DNA conformations and has been applied to study DNA bending of papillomavirus E2 protein binding sites and drug-DNA docking (Rohs et al., 2005a; Rohs et al., 2005b). A notable application is to investigate the intrinsic minor groove shapes of Exd-Scr heterodimer's specific *in vivo* site *fkh250* and its non-specific site $fkh250^{con}$ (Joshi et al., 2007). MC simulations suggested that *fkh250* have two narrow regions in the minor groove, whereas $fkh250^{con}$ has only one. Further studies showed that it is the extra narrow region in *fkh250* that attracted Scr's Arg3 and His-12 to bind and thereby made *fkh250* the specific site. Moreover, such feature of minor groove shape was also observed in the structures of Exd-Scr-*fkh250* and Exd-Scr-$fkh250^{con}$ complexes, suggesting that minor groove shape is intrinsic rather than induced by proteins and this intrinsic feature

can be detected by MC simulations. In addition, this MC method is able to predict helix twist observed from experimental structures (Rohs et al., 2009a). The success of MC method indicates that detailed description of DNA with a crude solvent model is capable in modeling sequence-dependent DNA conformations.

### 1.5.3 **All-atom simulations for protein-DNA interactions**

Both MD and MC simulations have been applied to study protein-DNA interactions. Valuable insights were obtained on the specific DNA-binding of p53, Lac repressor, and HU protein. Unlike free DNA structures, simulations for protein-DNA structures are usually based on CHARMM rather than AMBER.

The DNA site preference of tumor-suppressor protein p53 has been investigated by MD simulations (Pan and Nussinov, 2007, 2008). The canonical p53-binding site contains two decameric half-sites separated by base-pairs with variable length. MD simulations on two continued half-sites without insertion in the linker region demonstrated that bending of the DNA site increased p53-DNA interface contacts. Further MD study on p53 sites with variations in the central four base pairs of each half-site showed that DNA bending affected crucial p53 Arg280-guanine interactions, which in turn influence DNA flexibility. MD simulations have also been applied to zinc finger proteins to predict their specific binding sites (Seeliger et al., 2011) and the roles of interface water molecules at the zinc finger-DNA interface (Temiz and Camacho, 2009).

Several protein-DNA complexes have been studied by MC simulations, mainly through coarse-grained models. The binding of Lac repressor to its binding operator was found to be affected by sequence-dependent DNA conformations (Swigon et al., 2006).

Likewise, MC simulations demonstrated that the binding of nucleotide protein HU to its non-specific site were influenced by local DNA bending and untwisting (Czapla et al., 2008). MC simulation has also been applied to study the nucleosomes. The positively charged histone tail was revealed to be crucial for the conformational and dynamics properties of chromatin (Arya and Schlick, 2006). The relative orientation between linkers for a two-nucleosome array was found to induce twist and thus potentially control organization of chromatin fibers (Dobrovolskaia et al., 2010).

### 1.5.4 **Homology modeling on protein-DNA interactions**

The principle of homology modeling is based on the observation that sequence-similar proteins usually have similar structures (Petrey and Honig, 2005). Given the sequence of a target protein, a template selection is first performed to identity the protein structure that has close sequence similarity to the target. Based on the template structure, a homology model is built for the target, followed by a structural refinement process which optimizes the secondary structural elements and protein side-chain conformations. Homology models are helpful to infer protein functions, such as ligand-binding, protein-protein interactions, and protein-DNA interactions.

Homology modeling methods have been applied to study protein-DNA interactions and provided insights into the DNA-binding preferences of proteins. Siggers and Honig developed an all-atom homology modeling approach to predict the DNA-binding specificity of $C_2H_2$ zinc finger protein Zif268. They found that the prediction accuracy largely depended on the similarity of the interface docking geometry between template and target (Siggers and Honig, 2007). Ashworth et al. applied all-atom modeling software Rosetta to redesign endonuclease I-MsoI. Novel DNA cleavage specificity were

obtained and further confirmed experimentally (Ashworth et al., 2006). Alibes et al. used the modeling software FoldX to study DNA-binding specificity of a mammalian transcription factor PAX6, which play a crucial role in eye and neuron development. FoldX reproduced the experimental DNA-binding specificity and provided structural explanations for most of the known human mutations (Alibes et al., 2010). The threading-based method, DBD-Threader, was demonstrated to achieve considerably high sensitivity and precision in predicted DNA-binding domains and the associated DNA-binding residues by large benchmark tests (Gao and Skolnick, 2009). Morozov and Siggia applied homology modeling on a genome-wide scale. They used structural information to refine genomic binding sites for transcription factors and assigned potential binding sites to transcription factor families (Morozov and Siggia, 2007). All these investigations suggest that homology modeling method is a powerful tool to study the structural basis of protein-DNA specific interactions.

# Chapter 2. Background: Hox Proteins

## 2.1 What are Hox proteins?

Hox proteins are homeodomain-containing transcription factors that control body plan on the antero-posterior axis in both vertebrates and invertebrates (McGinnis and Krumlauf, 1992; Pearson et al., 2005). They are originally discovered in *Drosophila* where Hox mutations cause homeotic transformation of body segments (Lewis, 1978). The biological functions of Hox proteins and their implications in human diseases will be the topic for Section 2.2.

In *Drosophila*, there are eight Hox proteins: Labial (Lab), Proboscipedia (Pb), Deformed (Dfd), Sex combs reduced (Scr), Antennapedia (Antp), Ultrabithorax (Ubx), Abdominal-A (AbdA), and Abdominal-B (AbdB). All *Drosophila* Hox proteins have a conserved homeodomain composed of sixty amino acids. Hox homeodomains contain an unstructured "N-terminal" arm and a bundle of three alpha-helices, which are both employed to contact DNA (Gehring et al., 1994a). To understand the structure-to-function relationship of Hox proteins, many crystallographic and NMR studies have been carried out. The structural basis for their role in recognizing DNA and interacting with Exd will be introduced in Section 2.3.

Functioning as transcription factors, Hox proteins have high degree of *in vivo* DNA-binding specificity (Mann et al., 2009). Since all Hox proteins share the same homeodomain, it was speculated that Hox proteins would recognize and regulate the right subset of target genes by the DNA-interacting residues there. But, it turned out that

homeodomain is not the only determinant for specificity, raising the question of what the origin is for Hox specificity. Towards answering this question, a number of approaches were carried out from early days of biochemical experiments to nowadays of high-throughput measurements. The current knowledge on the origins of Hox specificity will be the topic for Section 2.4.

## 2.2 Biological functions of Hox proteins

### 2.2.1 **Define body plan on the anterior-posterior axis**

Hox genes are established during early *Drosophila* development by a combination of specific maternal, gap and pair-rule genes (St Johnston and Nusslein-Volhard, 1992). These upstream genes initiate and restrict transcriptional activation of Hox genes to a unique stripe of blastoderm cells. The regulation of Ubx, for example, has been shown to be directly mediated by a gap protein, hunchback, binding to an Ubx enhancer element (Qian et al., 1991).

After activation at the cellular blastoderm stage, the control of Hox proteins is maintained through a combination of auto-regulation, cross-regulation, and regulatory by Polycomb and Trithorax proteins. Auto-regulation was revealed by the persistent expression of Dfd. After initial activation in a blastoderm stripe, Dfd auto-regulates its transcription in epidermal cells through a positive feed-back loop (Regulski et al., 1991). Cross-regulation takes place when there are overlapped expressions of Hox genes. Usually, a more posteriorly expressed Hox gene represses, either partially or completely, a more anteriorly expressed Hox gene, referred as “posterior prevalence” rule (Lufkin et al., 1991). Ubx, for instance, normally establishes the PS6 identity. When it is

ubiquitously expressed, all thoracic and head segments are transformed towards PS6 identity. In contrast, this ubiquitous expression cannot transform abdominal segments towards PS6 identity. In addition to auto-regulation and cross-regulation, proteins of Polycomb group and Trithorax group also contribute to the persistent and refined expression of Hox genes. Both groups are chromatin regulatory proteins, but act antagonistically. Polycomb group function as repressors (Wedeen et al., 1986), whereas Trithorax group stabilize the expression of Hox genes (Shearn, 1989).

To define the embryo's body plan, Hox proteins act both as high-level executives to control other executive genes and as "blue collar" to mediate cell adhesion, cell cycle, cell death and cell movement. When Hox regulation takes place at the executive level, Hox proteins dictate the expression of transcription factors and morphogen signaling molecules, including Exd, Homothorax (Hth), Decapentaplegic (Dpp), Distalless (Dll), Nautilus, and Collier. Dpp, for example, is a secreted morphogen of the bone morphogenetic protein class and function as trigger for cell shape changes in gut required for normal visceral morphology. The expression of Dpp is activate by Hox protein Ubx and repressed by another Hox protein AbdA (Capovilla and Botas, 1998). Dll, as another executive genes regulated by Hox proteins, promotes appendage development. Its expression is directly repressed by both Ubx and AbdA, resulting in an absence of limbs from the abdomen (Vachon et al., 1992).

In order to sculpt morphology, Hox proteins need to regulate cell adhesion, cell cycle, cell death and cell migration. For cell adhesion, AbdA's mouse homolog HoxA13 has been implicated in mediating mesenchymal condensation in distal limb (Stadler et al., 2001). HoxA10, another mouse homolog of *Drosophila*'s AbdA, was shown to play a

role in controlling cell cycle (Thorsteinsdottir et al., 1997). Forced expression of HoxA10 induces premature differentiation of myelomonocytic cells into monocytes. Hox proteins have also been revealed to function as part of the cell replication machinery to coordinate cell growth and differentiation (Miotto and Graba, 2010). Cell death control is another way for Hox to regulate morphology. Dfd positively regulates the apoptosis-promoting gene *reaper* to maintain the segmental boundary between maxillary and mandibular segments of head (Lohmann et al., 2002). The ceh-13, a *C. elegans* Hox protein was found to function in the middle and posterior segments to control cell migration (Tihanyi et al., 2010).

### 2.2.2 **Cooperative DNA binding with cofactors**

It is well established that Hox proteins achieve specificity *in vivo* through cooperative DNA-binding with their cofactors (Mann, 1995; Mann and Carroll, 2002; Mann et al., 2009). In *Drosophila*, the known cofactors are Exd, Hth, and Engrailed. On one hand, cofactors can interact with Hox proteins to evoke their latent DNA-binding specificity (see Section 2.3 & 2.4). On the other hand, cofactors themselves have alternative splicing forms to introduce more Hox specificity.

Exd was first identified from a zygotic lethal mutations resulted in pattern defects in the first instar larva (Jurgens et al., 1984). Exd mutants were found to modify Ubx function even when Ubx was expressed ubiquitously throughout the embryo, suggesting Exd acts either in parallel to Ubx or downstream of Ubx (Peifer and Wieschaus, 1990). Further experiments showed that alterations in Hox gene expression were not observed in Exd mutants. This proved Exd's role as a cofactor for Hox protein. Through biochemical, *in vivo*, and structural studies, Exd was found to use its TALE motif, a three amino acid

loop extension, to make direct contacts with the YPWM motif of Hox proteins (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999). Although Hox protein AbdB does not have the YPWM motif, it kept a conserved Trp residue that can interact with TALE to enhance Hox-Exd interaction (Shen et al., 1997). Besides YPWM motif, a peptide immediately following Ubx homeodomain, termed UbdA, because of its similarity between Ubx and AbdA, also plays a role in Hox-Exd interaction (Merabet et al., 2007). In phenotypic suppression experiments, the UbdA motif of AbdA was revealed contributing to the competition of Exd-dependent DNA-binding (Noro et al., 2011).

Hth is another cofactor for Hox proteins. Like Exd, Hth is a member of the TALE family of homeodomain proteins. It shares significant homology with mouse protein Meis and Prep at both N-terminal and homeodomain (Rieckhof et al., 1997). Hth interacts with Exd via the N-terminal region. It is required for the nuclear localization of Exd and has been shown to interact with Exd in a DNA-independent way (Rieckhof et al., 1997). The expression of Hth controls Exd's ability to interact with Hox and influences Hox specificity. In mouse, Hth's homolog Prep1 was shown to bind to *R3* site together with Pbx-HoxB1 and increase the Pbx-HoxB1-dependent activation of *b1-ARE* (Berthelsen et al., 1998).

Engrailed, a homeodomain protein, has also been shown to be a Hox cofactor (Gebelein et al., 2004). It binds cooperatively with both Ubx and AbdA to the regulatory element of *Dll* gene. The input of Engrailed is required for repressing *Dll* in posterior compartments. Unlike Exd and Hth, which involve in either activation or repression with

Hox proteins on downstream genes, Engrailed can only function as a repressor in the Engrailed-Hox complex.

Alternative splicing of cofactors introduces more freedom for Hox specificity. Hth has two splicing forms: one with homeodomain and one without. The existence of both forms suggests that they may be used in different ways to achieve Hox specificity (Noro et al., 2006). Exd has only one isoform in *Drosophila*, but its vertebrate homolog Pbx produces multiple forms. Yeast two-hybrid assay suggests that these isoforms have distinct abilities to interact with Hth's vertebrate homologs, including Meis1, Meis2a, and Prep1 (Milech et al., 2001).

### 2.2.3 **Hox proteins in human diseases**

Hox proteins determine body development along the anterior-posterior axis. It is conceivable that Hox mutants would induce body malformation. Two syndromes, SPD and HFGS, have been found related to Hox mutations (Goodman, 2002). SPD is a dominantly-inherited limb malformation with a distinctive combination of syndactyly and polydactyly. This syndrome is caused by an expansion of polyalanine tract in the N-terminal region of HoxD13. Although the function of polyalanine tract is not well understood, it is possible that expansion of this region would disturb the interaction of HoxD13 with other proteins (Muragaki et al., 1996). HFGS is another dominantly-inherited disorder with distal limb abnormalities. Families with HFGS were identified with five types of mutations in HoxA13 (Goodman et al., 2000). The first three are nonsense mutations in either exon 1 or homeobox. The fourth is a polyalanine tract expansion similar to those in HoxD13 causing SPD. The fifth is an Asn to His mutation at a key DNA recognition position in HoxA13's homeodomain. This mutation induces

extremely short thumbs, reinforcing the importance on understanding DNA-binding specificity of Hox proteins.

Several types of cancer were caused by aberrant Hox gene expression (Abate-Shen, 2002; Shah and Sukumar, 2010). Overexpression of HoxA9 was observed in acute myeloid leukaemia by microarray analysis. Further studies found a fusion protein of HoxA9 and nucleoporin protein NUP98 from leukaemia cells resulted from rearrangement of chromosome 7 and 11 (Ghannam et al., 2004). Temporospatial deregulation of Hox proteins can induce cancer. HoxA5, normally expressed only in basal cells, was found to have expression in all oesophageal squamous carcinoma cells and promote tumor progression (Takahashi et al., 2007).

## 2.3 Structural studies of Hox-DNA complexes

### 2.3.1 **Monomeric Hox-DNA interactions**

Homeodomain is the major component used by Hox proteins to recognize specific DNA sites (Gehring et al., 1994a; Gehring et al., 1994b; Wolberger, 1996). It is a highly conserved DNA-binding domain composed of 60 amino acids. It consists of an N-terminal arm and three α-helices. The N-terminal arm is very flexible and only becomes ordered upon DNA-binding as a result from contacts with DNA minor groove. The three α-helices fold into a compact structure around a hydrophobic core, in which the third helix contacts DNA major groove.

Both NMR and crystallographic studies have been carried out to reveal detailed residue-DNA contacts for Hox protein Antp's homeodomain. NMR solved the first Antp-

DNA complex, where a full Antp homeodomain with a C39S mutation bound to Antp's specific site *BS2* (Otting et al., 1990). DNA minor groove was contacted by homeodomain's N-terminal arm, where Arg3 forms a salt bridge with phosphate group and Arg5 makes contacts with sugar moieties and bases. Another NMR structure of DNA in complex with a truncated Antp, which lacked residues 1-6, suggested a crucial role homeodomain's N-terminal arm to DNA-binding affinity. The truncated Antp still kept the same overall homeodomain structure, but its DNA-binding affinity was reduced by 10-fold (Qian et al., 1994). The third α-helix of homeodomain contacts DNA major groove in a relative orientation parallel along the groove, allowing an efficient way to stabilize homeodomain-DNA complex. From the two NMR structures, specific DNA recognitions were observed between DNA bases and three residues: Ile47, Gln50, and Met54.

The first X-ray structure of Hox's homeodomain-DNA complex was solved eight years after the report of the first NMR structure (Fraenkel and Pabo, 1998). In the X-ray structure, same full length Antp's homeodomain binding with same DNA were resolved at 2.4Å. The overall docking arrangement of homeodomain to DNA between X-ray and NMR structure are very similar. The crystallographic coordinates are generally consistent with protein-DNA NOEs from the NMR study, considering the flexibility of N-terminal arm and arginine and lysine side-chains. The major difference is on the conformation of Asn51. In the X-ray structure, Asn51 is in a well-defined predominant conformation making bidentate hydrogen bonds to an adenine base. The same Asn51 conformation has been observed in the crystallographic structures of other homeodomains, such as engrailed, paired, even-skipped, and yeast Mata1/α2. In contrast, due to the broad β

methylene and missing δ amino resonances, Asn51 in the NMR structure was suggested to make a fluctuating network of weak-bonding interactions with DNA. This disagreement could due to the magnetic environment around Asn51 is fluctuating, but the most populated and stable conformation is to form hydrogen bonds with DNA base. Nevertheless, both NMR and X-ray studies provide a detailed picture for illustrating the interaction between Hox protein's homeodomain and its DNA site.

### 2.3.2 **Exd-Hox-DNA interactions**

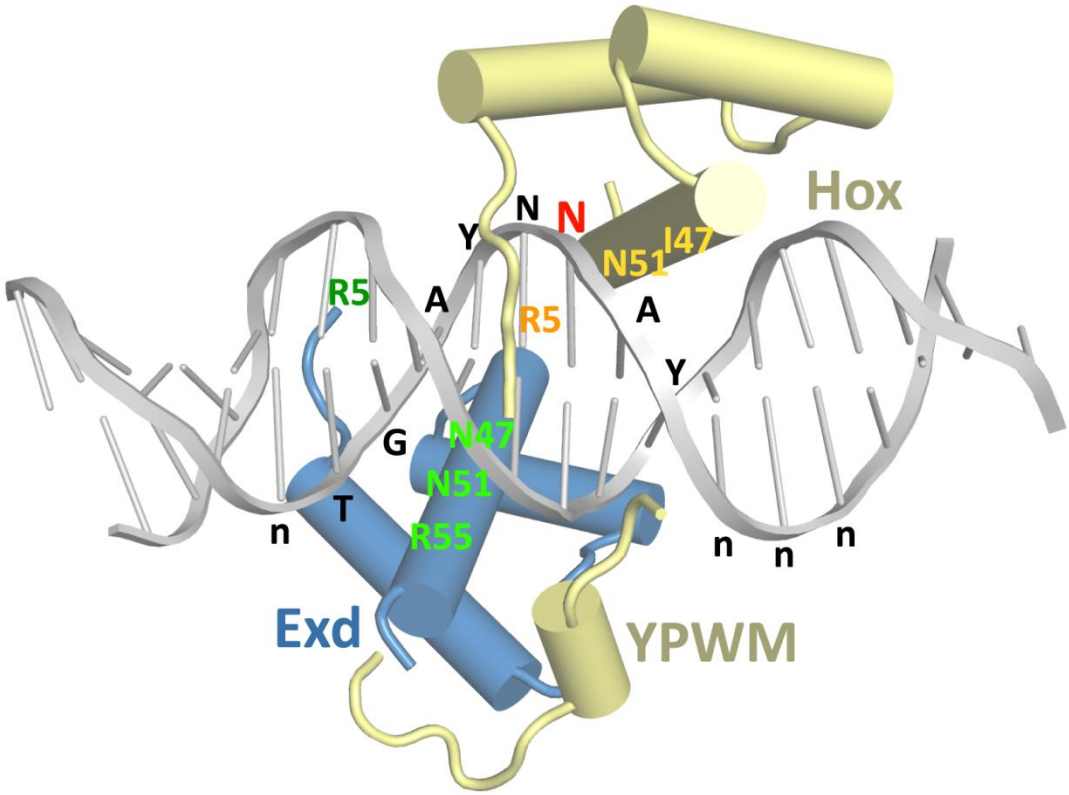


**Figure 2-1. A general representation of Exd-Hox heterodimer binding to DNA.**

The consensus 12-mer binding site is annotated along the DNA strand (N/n = any four types of bases; Y = T or C). Protein residues that have conserved interactions with DNA bases as observed from all five Exd/Pbx-Hox structures are labeled. This representation is made from the Scr-Exd-*fkh250* structure (PDB ID: 2r5z (Joshi et al., 2007)).

There are five structures have been solved for studying the cooperative binding of Hox and its cofactor Exd (or Pbx in mammals) to their DNA sites (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999). All these structure provide novel insights onto the Exd-Hox interaction and its role in DNA minor groove recognition (Figure 2-1).

The first Exd-Hox-DNA structure was solved in 1999 at 2.4Å (Passner et al., 1999). The homeodomains of Hox protein Ubx and Exd bind to DNA in a head-to-tail orientation. Three extra residues from Exd's homeodomain: Leu, Ser, and Asn (LSN motif), form part of a hydrophobic pocket to accommodate Ubx's YPWM motif. This hydrophobic binding leads to a loss of ~570$Å^2$ in solvent-accessible surface area for both Ubx and Exd. The Trp in Ubx's YPWM motif plays dominant role in the binding. It makes hydrophobic contact with the pocket and forms hydrogen bond with the Leu in LSN motif. The other three residues in the YPWM motif: Tyr, Pro, and Met, make less interactions with Exd, but they buttress the Trp in the hydrophobic pocket through stacking interactions. In addition to the YPWM-LSN interaction, the two DNA-recognition helices of Exd and Ubx also contribute to the cooperative binding. There is a loss of ~50$Å^2$ solvent-accessible surface area between the two helices since Ubx and Exd homeodomain are juxtaposed closely.

In the same year when the structure of Exd-Ubx-DNA was published, another structure, *Drosophila* Hox protein Lab's vertebrate ortholog HoxB1 and Exd's vertebrate ortholog Pbx1 were crystallized together with their DNA binding site (Piper et al., 1999). HoxB1-Pbx1 binds to the DNA in a very similar way as Ubx-Exd. It confirmed the interaction of Hox's YPWM motif (FDWM motif in HoxB1) with cofactor's LSN motif.

The third Hox-cofactor-DNA structure was solved in 2003 at a 1.9Å resolution (LaRonde-LeBlanc and Wolberger, 2003). It contains *Drosophila* AbdB's vertebrate ortholog HoxA9 binds to DNA together with Pbx1. In general, this structure is similar to the previous two in terms of Hox-cofactor interaction and DNA binding. A new feature uncovered by this structure is the minor groove recognition by Hox's N-terminal arm and the linker region. Arg2 on the N-terminal arm makes water-mediated hydrogen bonds with DNA bases in the minor groove. Leu-4, Ala-2, and Lys3 form either direct or water-mediated hydrogen bonds with minor groove backbone phosphates. Although the DNA site in this structure is not an *in vivo* binding site, the contacts between N-terminal arm and minor groove observed from HoxA9 suggest the potential generality of DNA minor groove recognition by Hox-cofactor complexes.

The functional importance of minor groove recognition was uncovered by a pair of Exd-Scr-DNA structures solved in 2007 (PDB ID: 2r5z & 2r5y solved at resolution of 2.6 Å) (Joshi et al., 2007). Both structures contain the same heterodimer Exd-Scr, but with different DNA-binding sties: one is Scr's specific site *fkh250*, which only binds to Scr but not to other Hox proteins, the other is Scr's non-specific site $fkh250^{con}$, which can also be recognized by other Hox proteins, such as Ubx and AbdA. The primary goal in solving these two structures is to identify the structural determinant for Scr-*fkh250* specificity. It turns out that the specificity origins from the minor groove recognition. In the Exd-Scr-*fkh250* structure, two residues from Scr's linker region and N-terminal arm: His-12 and Arg3 were found binding to *fkh250*'s minor groove (Figure 2-2). By contrast, in the Exd-Scr-$fkh250^{con}$ structure, no such binding was observed. Further analysis shows that *fkh250* has an extra narrow region in the minor groove that can induce enhanced

electrostatic potential, which attract His-12 and Arg3 to bind. In the non-specific site *fkh250*$^{con}$, on the other hand, an adenine is replaced by a thymine at position and therefore the second A-tract is removed. The corresponding minor groove region becomes wide and the electrostatic potential is no longer that favorite to attract His-12 and Arg3.

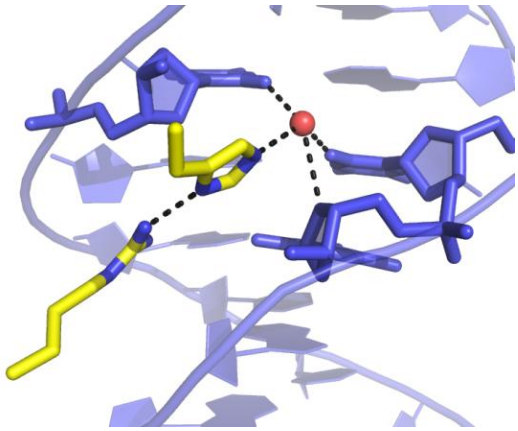

**Figure 2-2. Scr's Arg3 and His-12 bind to narrow minor groove.**

Arg3 and His-12 are shown in sticks. Hydrogen bonds are represented by dashed lines and water molecule is denoted by red sphere. The PDB ID for this structure is 2r5z.

All five Hox-cofactor-DNA structures show the same head-to-tail DNA-binding orientations and conserved major groove recognitions for Hox protein and its cofactor. The cooperativity is from the interaction of Hox's YPWM motif and the hydrophobic pocket mainly created by the extra three residues in Exd/Pbx's homeodomain. This protein-protein interaction not only stabilizes the cooperative DNA-binding of Hox and its cofactor, it also helps to steer Hox's linker region and its N-terminal arm into DNA minor groove so that Hox can specifically recognize its DNA site.

## 2.4 DNA-binding specificity of Hox proteins

### 2.4.1 **Three levels of Hox DNA-binding specificity**

By considering the target genes, Hox specificity can be classified into three levels: “paralog-specific”, “semi-paralog-specific”, and “general” (Mann et al., 2009). The “paralog-specific” category refers to genes that are only regulated by a single Hox paralog. For example, *fkh250* is a 37-base-pair element directly regulated by Scr for the development of salivary gland. Misexpression of Scr throughout *Drosophila* embryo will ectopically activate *fkh250*-lacZ, even in presence of more posterior Hox (Bradley et al., 2001).

“Semi-paralog-specific” target genes are those shared by a subset of Hox proteins. *Distalless* (*Dll*), for instance, is repressed by abdominal Hox proteins Ubx, AbdA and AbdB, thereby limiting the leg development only to the thoracic segment. Furthermore, Ubx and AbdA were found to work through common binding sites to regulate *Dll* expression (Gebelein et al., 2002).

A third category of Hox-specific target genes are those shared by most of Hox proteins. The *Drosophila* head-promoting gene, *optix*, for example, is repressed by the trunk Hox proteins (Dfd and Scr) and abdominal Hox proteins (Antp, Ubx, and AbdA), but is activated by more anterior Hox proteins (Lab and Pb) (Coiffier et al., 2008).

In addition to the three levels of Hox specificity, developmental context is another issue need to be considered. For some tissues, only one Hox protein is expressed, whereas all the other seven are not. For instance, only Hox protein Ubx is expressed in the developing haltere cells (Crickmore and Mann, 2008) and it targets a variety of genes at different stages of appendage morphogenesis (Pavlopoulos and Akam, 2011). The cis-

regulatory elements for Ubx are not required to be highly specific because all the other Hox proteins never exist within the same tissue. When expressing other Hox proteins in the wing, Ubx-like regulation will be observed, confirming the importance of developmental context to Hox specificity (Casares et al., 1996).

### 2.4.2 **DNA-binding specificity of Hox homeodomains**

Early biochemical studies have established the DNA-binding specificity for some Hox proteins. Ekker et al. found that homeodomains of Dfd and Ubx both prefer to bind to the core motif 5'-TAAT-3' (Ekker et al., 1992). The different specificity for these two Hox proteins occurs at the flanking sequences. On the 5' flanking region, Ubx prefers T or C, whereas Dfd only prefers T. On the 3' flanking region, Ubx prefers G > TpG > ApCpC, while Dfd likes G > TpA > GpApC. Later work extent the study to another two Hox proteins, Antp and AbdB (Ekker et al., 1994). Similar to Dfd and Ubx, Antp also prefer a core motif of 5'-TAAT-3'. By contrast, AbdB likes to bind to a different core motif 5'-TTAT-3'. On the flanking sequences, Antp displays nearly identical preferences as Ubx.

Recently, two high-throughput approaches have been carried out to identify DNA-binding specificities for nearly all homeodomains in *Drosophila* (Noyes et al., 2008a) and mouse (Berger et al., 2008). One approach used a bacterial one-hybrid (B1H) system to define DNA site preferences for 84 homeodomains from *Drosophila*, while the other study characterized the DNA-binding specificities of 168 mouse homeodomain by protein binding microarrays (PBM). Both works found that Hox homeodomain prefer to bind "AT"-rich sequences, and the B1H approaches confirmed previous biochemical studies that, the so-called "Antennapedia group", which contains all eight *Drosophila*

Hox proteins except AbdB, prefer to bind 5'-TAAT[t/g][a/g]-3'. However, there are more than 80 thousand copies of TAATTA or TAATGA in the *Drosophila* genome and these sites are not only bound with Hox proteins, but also to other non-Hox homeodomains (Mann et al., 2009). Therefore, the 5'-TAAT[t/g][a/g]-3' sites are not sufficient to explain the distinct functions of *Drosophila* Hox proteins *in vivo*.

### 2.4.3 **Cofactors evoke the latent DNA-binding specificity of Hox proteins**

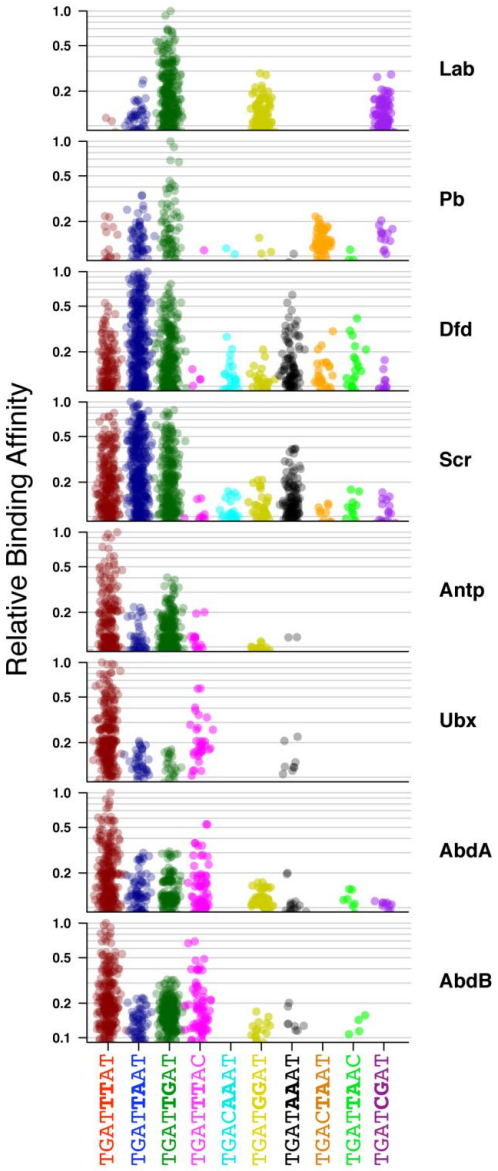


**Figure 2-3. DNA-binding specificity for Exd-Hox complexes.**

Strip charts (with arbitrary horizontal displacement) showing the distribution of relative affinities (y-axis) across all 12-mers (x-axis) for eight Exd-Hox measured by SELEX-seq. Figure adapted from a published paper (Slattery et al., 2011b).

DNA-binding specificity of Hox proteins is modulated by their cofactors. Unlike Hox monomers, Pbx-Hox displays a bipartite preference on the sequence 5'-ATGATTNATNN-3' (Chang et al., 1996). The cofactor Pbx contacts the 5' half ATGAT and Hox protein bind to the 3' half TNATNN. Moreover, Hox proteins show stepwise differences in their DNA-binding specificity on the 3' half when they bind DNA together with Pbx: an increasing preference of $G_7$ towards HoxB1, an increasing preference of $T_7$ towards HoxB9 and HoxA10, and a shared preference of $A_7$ from HoxB3 to HoxB9.

Recently, a combined Systematic Evolution of Ligands by Exponential Enrichment with massively parallel sequencing (SELEX-seq) was employed to characterize DNA-binding sites for all eight full-length *Drosophila* Hox proteins in complex with their cofactors: full-length Exd and the "Homothorax-Meis" domain of Hth (simply referred them as "Exd") (Slattery et al., 2011b). A characteristic affinity "fingerprint" was found across eight Exd-Hox complexes (Figure 2-3). For instance, Exd-Lab and Exd-Pb do not bind well to 5'-TGATTTAT-3', while only Exd-Ubx fails to bind to 5'-TGATGGAT-3'. Based on the specificity profiles, the eight Hox proteins can be categorized into three classes, where class 1 contains Lab and Pb; class 2 consists of Dfd and Scr; and class 3 encompasses Antp, Ubx, AbdA, and AbdB. Hox proteins within the same class have similar *in vitro* DNA-binding preference.

This SELEX-seq experiment demonstrates unique DNA-binding preferences of Hox upon heterodimerization with Exd. Two pairwise comparisons of monomeric Hox-binding preferences (Scr vs. Lab and Scr vs. Ubx) show that the general tendency for all three Hox proteins select sequences containing 5'-TAAT-3' (Figure 2-3 A and B). By contrast, when in complex with Exd, the same Hox proteins display a high degree of

specificity. The 5'-TGATTAAT-3' and 5'-TGATTGAT-3' bound more strongly to Exd-Scr than to Exd-Ubx (Figure 2-3 D). Conversely, the 5'-TGATTTAC-3' site has higher affinity to Exd-Ubx than to Exd-Scr (Figure 2-3 D). Similarly, in presence of Exd, the specificities of Scr and Lab are distinguishable (Figure 2-3 E), while the corresponding monomeric specificities are largely overlapped (Figure 2-3 B). Comparisons between AbdB and Ubx reveal another type of Exd-dependent change in DNA-binding specificity. Monomers of AbdB and Ubx have both common and unique binding site preferences (Figure 2-3 C), whereas Exd-AbdB and Exd-Ubx share very similar specificities: both prefer 5'-TGATTTAT-3' and 5'-TGATTTAC-3' (Figure 2-3 F). Therefore, it is the heterodimerization with Exd converge the specificities of these two Hox proteins.

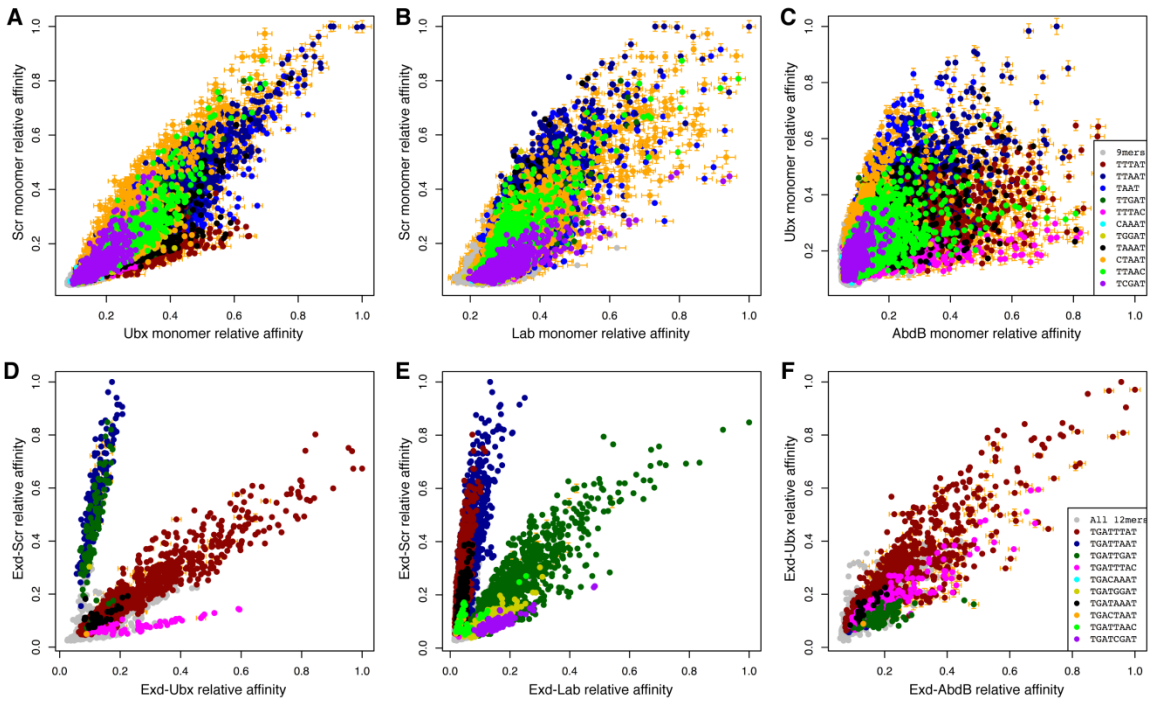


**Figure 2-4. Heterodimerization with Exd induces novel binding specificities.**

Comparisons of the specificity for monomeric Hox proteins on all 9-mers (A-C) and for Exd-Hox dimers on all 12-mers (D-F). Figure adapted from a published paper (Slattery et al., 2011b).

# Chapter 3. DNA Minor Groove Shape Is a Structural Determinant for the Specificity of All Eight Drosophila Hox Proteins

## 3.1 Introduction

Gene regulatory information is encoded in genomic DNA sequences and interpreted by factors that bind to these sequences. Critical players in this decoding mechanism are proteins that recognize DNA in a sequence-dependent manner. Although the *in vitro* binding properties of transcription factors (TFs) have been studied for many years, it has proven notoriously difficult to predict *in vivo* genomic binding from *in vitro* sequence specificity. Whether or not a predicted binding site is occupied *in vivo* depends strongly on sequence and chromatin context as well as cell type (Gaulton et al., 2010; Guertin and Lis, 2010; Kaplan et al., 2011; Natoli, 2010).

What makes *in vivo* binding more specific than *in vitro* binding? One possible answer has its root in the combinatorial nature of gene regulation. Unlike individual TFs, complexes of interacting factors bind cooperatively to genomic regions that contain a favorable configuration of binding sites (Johnson, 1995; Panne, 2008). These mechanisms, however, are unlikely to be sufficient to account for the TF specificities observed *in vivo*. In particular, confounding the issue of specificity is that most TFs are members of protein families that have very similar DNA-binding domains with similar recognition properties. Despite overlapping binding specificities, these factors carry out distinct functions *in vivo* (Alexander et al., 2009; Cao et al., 2010; Kassouf et al., 2010;

Naiche et al., 2005; Pearson et al., 2005). The fundamental question of how they recognize distinct binding sites and regulate unique sets of target genes *in vivo* remains unsolved.

Although members of the same TF family typically have very similar DNA-binding domains, these domains are rarely identical. This raises the possibility that small differences in protein sequence could lead to significant differences in binding specificity. On one hand, TF recruits different co-activators or co-repressors to diversify specificity (Gebelein et al., 2004; Hersh and Carroll, 2005; Joshi et al., 2010; Li and McGinnis, 1999; Taghli-Lamallem et al., 2007). On the other hand, differences in the amino acid sequences of TFs within the same structural family may only impact DNA recognition when these factors bind with cofactors.

The eight Hox paralogs in *Drosophila*, which execute distinct functions *in vivo*, each have recognizable orthologs in both vertebrates and invertebrates. A Systematic Evolution of Ligands by Exponential Enrichment with massive parallel sequencing (SELEX-seq) approach was employed to demonstrate that complex formation between *Drosophila* Hox proteins and Exd uncovers latent DNA-binding specificities that are only revealed upon heterodimerization. By analyzing the enrichment of oligonucleotides through several rounds of selection, all eight Exd-Hox heterodimers, regardless of the Hox protein, were found to prefer to bind the sequence 5'-GAYNNAY-3' (where Y=T or C) and that the familiar preference of Hox proteins for 5'-TAAT-3' sequences no longer dominates. Different Exd-Hox heterodimers exhibit strong preferences for distinct subsets of this generalized binding site, leading to a unique binding "fingerprint" for each Exd-Hox complex. Strikingly, these fingerprints naturally cluster into three groups,

which correspond to the Hox expression domains along the anterior-posterior axis. More generally, these results suggest that members of transcription factor families achieve specificity in part by forming complexes that modify their DNA recognition properties in precise ways. This tuning involves the combined use of sequence-dependent properties in both the major and minor grooves.

Here we investigate the role of DNA minor groove shape in determining Hox specificity. Minor groove shape has been revealed before as a determinant for the specificity of a Hox protein Scr (Joshi et al., 2007) and this type of shape recognition is widely used for protein-DNA interaction (Rohs et al., 2009b). We interested in the question whether DNA minor groove shape is also a structural determinant for all eight *Drosophila* Hox proteins and what type of shape each Hox protein preferred. Using a Monte Carlo simulation method (Rohs et al., 2005b; Sklenar et al., 2006) and a knowledge-based computational method, we predicted the minor groove widths for Hox-binding sites obtained from the SELEX-seq experiment and made comparisons to study their similarities and differences. In addition to the study on DNA shape, we looked for paralog-specific residues that correlate with Exd-Hox specificity. Through examining their interactions with DNA from structures, we propose potential roles of these residues in Hox-DNA specific recognition.

## 3.2 Results

### 3.2.1 **Highest-affinity binding sties for Exd-Scr and Exd-Ubx have distinct shape**

Given the sequences obtained from SELEX-seq experiment, we examined the extent to which DNA minor groove width contributes to binding site preferences as

observed previously for Exd-Scr complexes (Joshi et al., 2007) and for other DNA binding proteins (Rohs et al., 2009). To address this question, we used all-atom Monte Carlo (MC) simulations (Joshi et al., 2007; Rohs et al., 2005b) to predict the width of the minor grooves of Hox-binding sites identified by SELEX-seq. We first carried out calculations on the ten DNA sequences with the highest binding affinities for Exd-Scr, which all contain a blue (5'-TGATTAAT-3') binding site. All ten sequences have a similar shape, with two narrow regions in the core, at $A_4T_5$ and $A_8T_9$ (numbering is based on the 12 bp site) (Figure 3-1 A). This double-minimum pattern of minor groove width is similar to the minor groove topology in an Exd-Scr crystal structure (Joshi et al., 2007) (Figure 3-1 A). One difference between several of the predicted versus crystal DNA structures is that the narrow region at $A_4T_5$ of the SELEX-seq sites extends over a greater number of base pairs due to a short A-tract $A_{-1}A_1T_2$ in the Exd flank. In the crystal structure, this A-tract is replaced by $T_{-1}A_1A_2$, which contains a TpA base pair step that tends to widen the minor groove.

We next used MC simulations to predict the shape of the ten sequences that have the highest binding affinity for Exd-Ubx, which preferentially binds the red (5'-TGATTTAT-3') core motif. Unlike the top-ranked Exd-Scr binding sites, all ten Exd-Ubx-preferred sequences have a narrow minor groove in the $A_4T_5$ region, and nine out of ten have a wider minor groove at $A_8T_9$ (Figure 3-1 B). Again, this pattern mirrors that observed in an X-ray structure of Exd-Ubx bound to DNA containing the core sequence of the red motif (Passner et al., 1999; Rohs et al., 2009b). Together, these simulations reveal unique shapes for the Exd-Scr preferred and Exd-Ubx preferred binding sites.

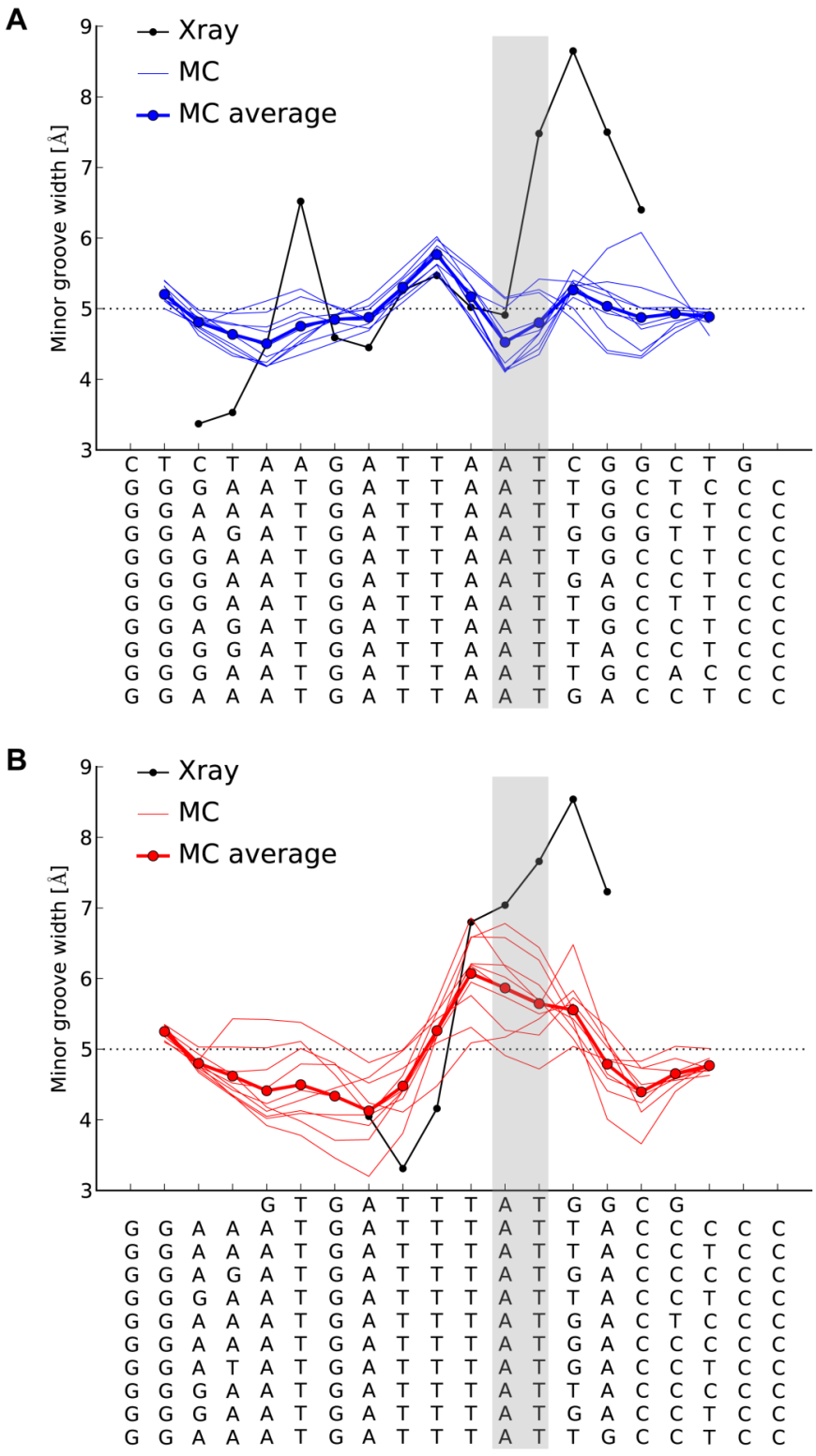


**Figure 3-1. Predicted minor groove widths of Exd-Scr and Exd-Ubx binding sites.**

(A and B) Shown are MC calculations of minor groove width on selected binding sites for Exd-Scr (A) and Exd-Ubx (B). The groove widths of DNA from crystal structures (black) of Exd-Hox-DNA ternary complexes (Joshi et al., 2007; Passner et al., 1999) are compared with the widths predicted by MC simulations for the ten highest affinity binding sites (thin blue lines in (A) and thin red lines in (B)). The average groove widths of the top ten sites in each group (thick blue line in (A) and thick red line in (B)) are also shown. The sequences for DNA-binding sites from crystal structures (top) and sequences for the ten SELEX-seq DNA sites are listed below the x-axis. The gray shading highlights the positions of $A_8T_9$ in the 12 base-pair binding site.

### 3.2.2 **DNA shape contributes to Exd-Hox dimer preferences**

The observation on the distinct DNA shape preference of Exd-Scr and Exd-Ubx intrigued us with the question that what the role of DNA shape is for all eight *Drosophila* Hox proteins. To answer this question, we extended our structural analysis to the top ten high affinity sites of two most favored motifs selected by all eight Exd-Hox complexes. Although these predictions do not reveal a distinct shape for each of the eight Exd-Hox complexes, they do suggest that Hox proteins with a closer functional relationship prefer to bind DNA sites with a related shape. Sequences preferred by Lab, Pb, Dfd, and Scr, which generally define anterior identities, all have narrow minor grooves or local minima in minor groove width at $A_8T_9$ (Figure 3-2 A). In contrast, DNA sequences preferred by Ubx, AbdA, AbdB, and Antp, which generally dictate posterior identities, all have wide minor grooves at $A_8T_9$ or $A_8C_9$ (Figure 3-2 B). The resulting trends in predicted structures suggest that there are two major groups of sequences with respect to DNA shape: those preferred by the four anterior Hox proteins and those preferred by the four posterior Hox proteins. The results also suggest that different DNA sequences, such as the green and blue sites preferred by Scr, can have similar shapes.

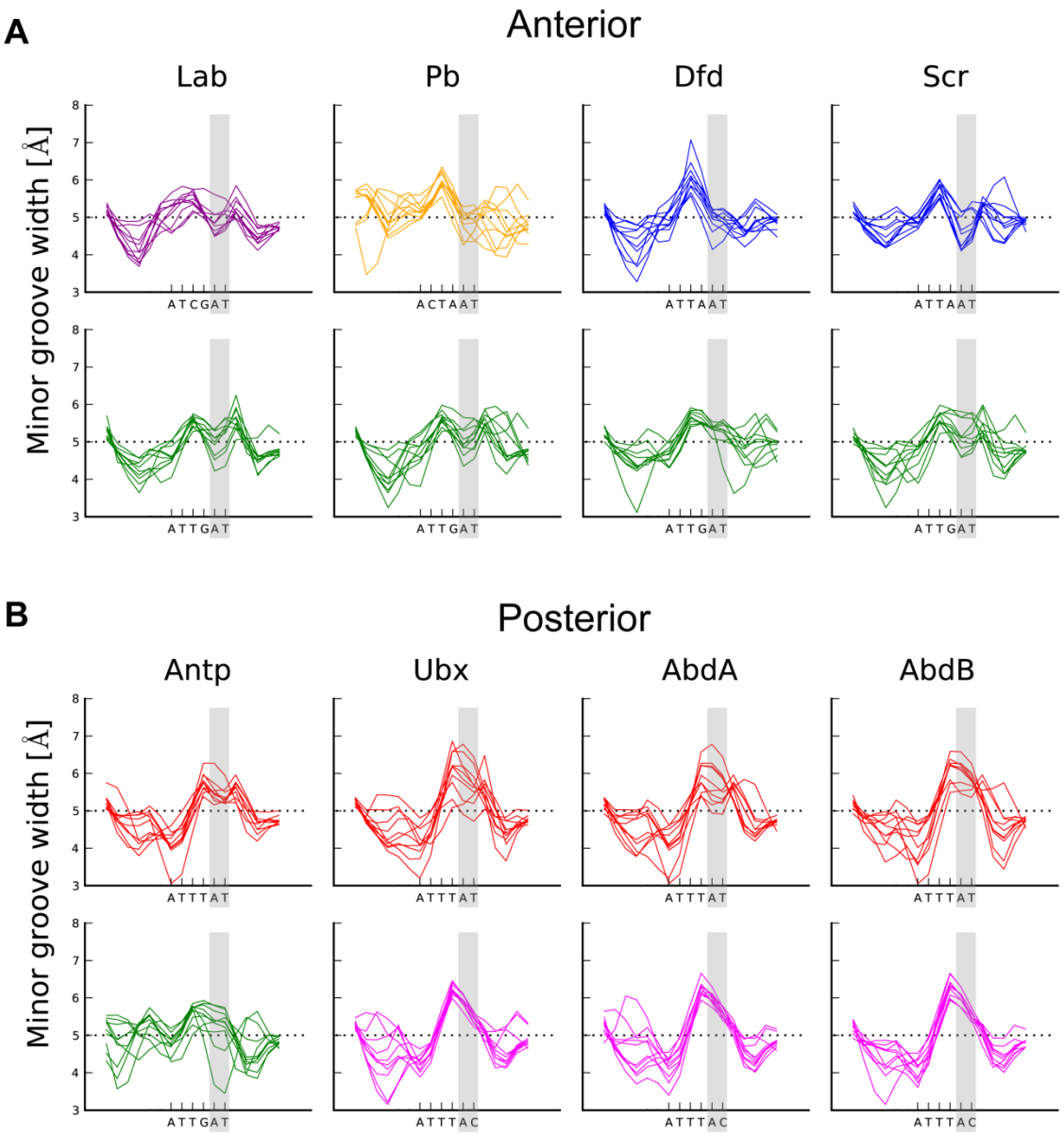


**Figure 3-2. Predicted minor groove widths of Exd-Hox binding sites.**

(A and B) MC predictions of minor groove width on selected binding sites for Exd in complex with anterior Hox (A) and posterior Hox (B). Individual MC prediction is represented in thin line. The core motif of DNA-binding site in each graph is listed below the x-axis, and the Hox protein identity is indicated on the top of each column. Gray shading highlights $A_8Y_9$.

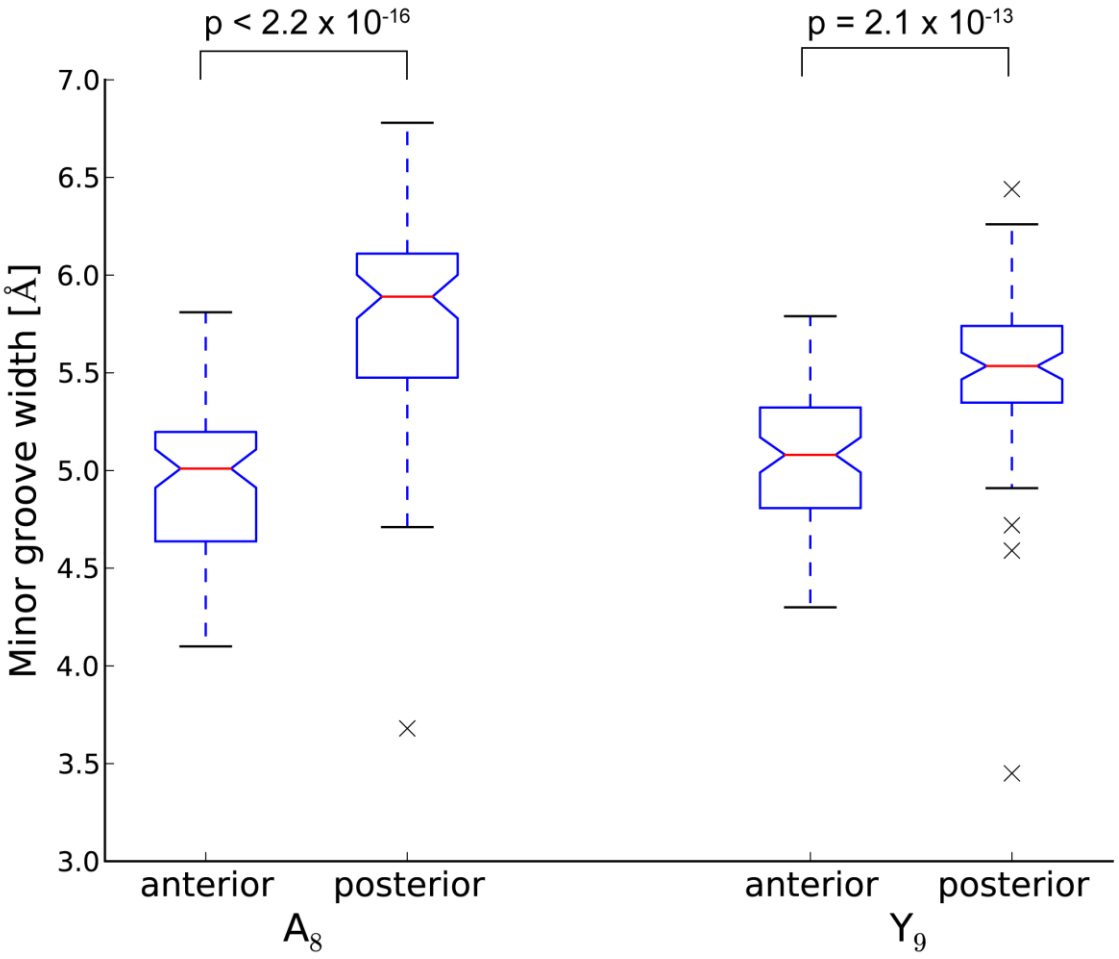


**Figure 3-3. The preference of minor groove shape between anterior Hox and posterior Hox is significantly different.**

Box plot compares the predicted minor groove widths at $A_8$ and $Y_9$ for motifs favored by anterior Hox and posterior Hox as shown in Figure 3-2. The p-values calculated from Mann-Whitney U tests are denoted on top of each comparison.

To quantitatively evaluate the difference between anterior shape and posterior shape, we performed a Mann-Whitney U test on the predicted minor groove widths. At $A_8$, minor groove widths for anterior Hox binding sites ranges from 4.0Å to 6.0Å. This distribution is relatively lower than minor groove widths for posterior Hox binding sites,

which mainly ranges from 4.5Å to 7.0Å (Figure 3-3). Mann-Whitney U test on the two groups of minor groove widths gave a p-value lower than $2.2*10^{-16}$, suggesting that the anterior shape and posterior shape at $A_8$ are significantly different. Similarly, at $Y_9$, anterior Hox binding sites also have minor groove widths distributed at relatively lower values than posterior Hox binding sites (Figure 3-3). Mann-Whitney U test gave a p-value equal to $2.1*10^{-13}$, reinforcing the significant difference of minor groove widths between the two groups. The quantitative comparison on minor groove widths at $A_8$ and $Y_9$ confirmed significantly different preferences of DNA shape for anterior Hox proteins and posterior Hox proteins.

To test whether the DNA shape preferences also holds true for a large number of sequences, we employed a high-throughput approach that predicts minor groove width based on the average conformations of tetra- and penta-nucleotides derived from >1600 MC simulations. Using this approach, we predicted the minor groove widths of all SELEX-seq sequences that have relative affinities above 0.1. Consistent with the previous shape analysis, all sequences, independent of Hox protein, had a minimum near $A_4$ (Figure 3-4). By contrast, binding sites preferred by anterior Hox proteins had on average narrow minor grooves at $A_8Y_9$, while those preferred by posterior Hox proteins had on average wide minor grooves at $A_8T_9$ or $A_8C_9$ (Figure 3-4). Furthermore, clustering based on the Euclidian distance between shape profiles along the central 5'-AYNNAY-3' motif was found to be compatible with the collinear ordering of the Hox proteins from anterior to posterior (Figure 3-4). This result shows that DNA shape preference exists for a large number of sequences and it is remarkable as it stems only from the predicted minor groove shapes of the SELEX-seq-derived binding sites.

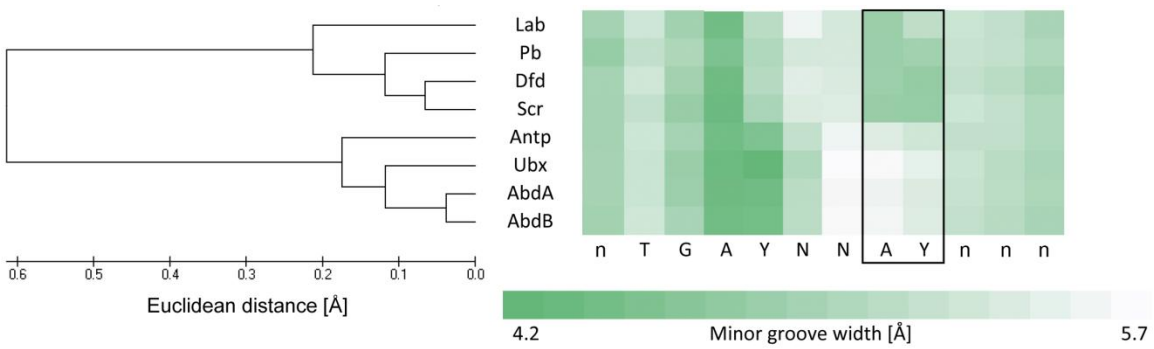


**Figure 3-4. Clustering of Hox proteins based on their preferences of DNA shape.**

(left) Dendrogram comparing minor groove shape for Exd-Hox binding sites based on Euclidean distances between average minor groove widths of core motif 5'-AYNNAY-3'. (right) Heatmap representing the average minor groove widths of all sequences above a relative binding affinity threshold of 0.1 for eight Exd-Hox heterodimers. Dark green represents narrow minor grooves and white denotes wider minor grooves. Figure adapted from a published paper (Slattery et al., 2011b).

### 3.2.3 **Conservation of Hox protein sequences correlates with DNA-binding specificity**

Although homeodomain and YPWM motif in Hox proteins are conserved, there still exists paralog-specific sequence, which is another major source beside intrinsic DNA shape, to determine Hox specificity (Figure 3-5). Given the diverse DNA preferences of Hox proteins measured from SELEX-seq experiments, we want to find out i) if there is paralog-specific residue that correlates with DNA-binding specificity, ii) what their roles are in terms of three-dimensional structures, and iii) what their contributions are to Hox specificity. Based on the multiple sequence alignment of arthropod Hox proteins, we found several paralog-specific residues (Figure 3-5 A). Tyr8, Leu14, and Arg43 correlate with Hox proteins' preference to the red core motif (5'-ATTTAT-3').

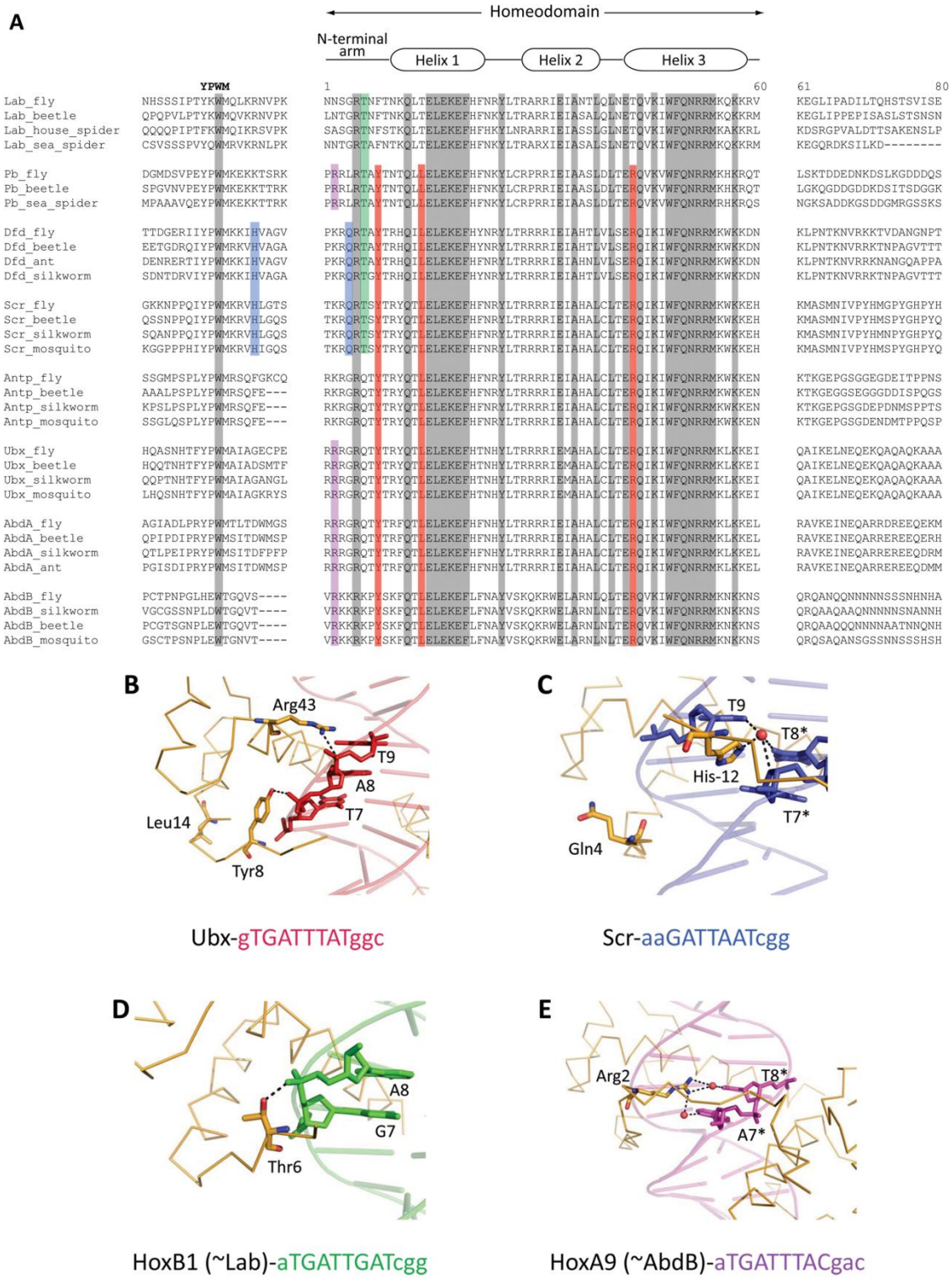


**Figure 3-5. Paralog-specific Hox residues correlate with DNA-binding specificity.**

(A) Alignments of sequences surrounding homeodomains and YPWM motifs of Hox proteins from arthropods. (B-E) Existing PBC-Hox X-ray crystal structures (PDB IDs: (B) 1b8i (Passner et al., 1999); (C) 2r5z (Joshi et al., 2007); (D) 1b72 (Piper et al., 1999); (E) 1puf (LaRonde-LeBlanc and Wolberger, 2003). The name of the Hox protein and its 12-mer DNA-binding sites are listed below each panel. Partially conserved Hox residues that correlate with DNA-binding specificity are highlighted by the colors in (A) and are represented in sticks in (B)-(E). Nucleotides that form direct or water-mediated hydrogen bonds with these residues are labeled and represented in sticks as well. The numbering of labeled nucleotides is based on the 12-mer binding site. Nucleotides on the complementary strand are labeled with stars. Hydrogen bonds are represented as dashed lines and water molecules are shown in red spheres.

The structure of Exd-Ubx-DNA showed that both Tyr8 and Arg43 make hydrogen bonds with phosphate groups at $T_7A_8T_9$ (Figure 3-5 B), while Leu14 positions away from DNA (Figure 3-5 B), possibly involving in interacting with other proteins *in vivo*. His-12 and Gln4 correlate with Hox specificity to the blue motif (5'-ATTAAT-3'). In the structures of Exd-Scr-*fkh250*, His-12 forms water-mediated hydrogen bond with DNA bases in minor groove and Gln4 is part of a RQR motif, which specifically recognizes minor groove with two narrow regions as suggested before (Figure 3-5 C). Arg2 and Thr6 correlate with Hox specificity to the green motif (5'-ATTGAT-3') and magenta motif (5'-ATTTAC-3'), respectively. X-ray structures indicate that Thr6 makes hydrogen bond with phosphate group between $G_7$ and $A_8$, and Arg2 recognizes DNA minor groove through water-mediated hydrogen bonds. All these paralog-specific residues suggest potential roles for Hox specificity in terms of three-dimensional structures.

## 3.3 Discussion

We have demonstrated that DNA minor groove shape contributes to Exd-Hox binding site preferences. This specificity origins from the intrinsic properties of DNA and is recognized when Hox proteins bind DNA together with their cofactor Exd. As such, these results provide a precedent for how interactions between DNA-binding proteins can result in emergent recognition properties that are not exhibited by either factor on their own. Clustering of Hox proteins' DNA shape preferences reproduced the ordering of their location along chromosome and the body segments where Hox proteins expressed along the anterior-posterior axis. This striking observation connects DNA minor groove width preferences with *Drosophila* morphogenesis.

### 3.3.1 **A single cofactor reveals latent DNA-binding specificities that distinguish members of the same transcription factor family**

As monomers, the eight Hox proteins in *Drosophila* recognize an overlapping set of AT-rich hexameric binding sites (Mann et al., 2009; Noyes et al., 2008a). In presence of Exd, however, we find that Hox's DNA-binding preferences become more focused and specific. These findings raise an important question: how can the same cofactor elicit unique specificities for eight closely related homeodomain proteins? We propose that the additional specificity information that is used to distinguish Exd-Hox binding preferences comes from the Hox protein, but that this information cannot be used effectively without Exd. In other words, Exd unlocks latent specificities that are present within the Hox protein sequences. It is plausible that other protein families also use an analogous mechanism to fine tune their DNA-binding specificities. For example, Runt domain proteins bind DNA with a higher degree of specificity when partnered with the cofactor CBF (core-binding factor) and different combinations of bHLH proteins appear to prefer

different E-box sequences (Bartfeld et al., 2002; Grove et al., 2009; Tahirov et al., 2001). We further speculate that novel DNA-binding specificities may not only arise from pairs of transcription factors; higher-order specificities may emerge as a consequence of the assembly of multi-protein-DNA complexes.

How might this work in molecular terms? For Hox proteins, one source of latent specificity information is likely to be in the N-terminal arms of their homeodomains and neighboring linker sequences. By binding to the 'YPWM' motif, which is located N-terminal to Hox homeodomains (Figure 2-1), Exd limits the structural freedom of this portion of Hox protein. For Scr, as seen in X-ray crystal structures, the YPWM-Exd interaction positions this region of Hox protein so that it can bind to the minor groove, primarily via three basic residues: two Arg (Arg3 and Arg5 of the homeodomain) and a His (His-12) (Joshi et al., 2007). Importantly, several residues in Scr's N-terminal arm and linker region are conserved in a paralog-specific manner and are important for executing Scr-specific functions *in vivo* (Joshi et al., 2007). Some of these residues correlate with the binding specificities identified here. For example, both Dfd and Scr, but none of the other Hox proteins, have His at position -12 (numbering is relative to the start of the homeodomain; Figure 3-5 A). Further, only Dfd and Scr have the N-terminal arm motif 'RQR' (where the first Arg is Arg3; Figure 3-5 A). Although other Hox proteins have an Arg at position 3, the adjacent Gln is unique to Dfd and Scr. The Gln4 is required for optimal binding, perhaps by favoring a conformation in which both Arg3 and Arg5 can insert into the minor groove (Joshi et al., 2007). Based on these correlations, we suggest that the RQR motif contributes to the preference that Dfd and Scr exhibit in the SELEX-seq experiments (Figure 2-3). Additional correlations between Hox protein

sequences and SELEX-seq binding site preferences are also apparent (Figure 3-5). For example, Ubx, AbdA, and AbdB have an Arg at position 2 of the homeodomain. In a crystal structure of the vertebrate AbdB ortholog HoxA9 bound to DNA in complex with Pbx, this Arg makes multiple water-mediated hydrogen bonds in the minor groove of a magenta binding site (Figure 3-5 E) (LaRonde-LeBlanc and Wolberger, 2003; Mann et al., 2009). Together, these observations argue that seemingly small differences in protein sequence between Hox proteins are exploited by Exd to help achieving DNA-binding specificity.

Although Hox homeodomain and linker sequences are likely to be important determinants of the DNA-binding specificities observed in SELEX-seq experiments, they are unlikely to account for all of the differences we observe between Exd-Hox complexes. One reason is that the proteins used in all of the existing crystal structures are primarily limited to the DNA-binding domains, while proteins used in the SELEX-seq experiments are significantly longer and in many cases, close to full-length. *In vitro*, the protein fragments used in the crystal structures bind to their binding sites with significantly less cooperativity than full-length proteins, suggesting that additional interactions are likely to exist in the native complexes. Additional structural studies using full-length proteins and alternative binding sites will be needed to fully understand the specificities revealed by SELEX-seq experiments.

### 3.3.2 **The role of DNA shape in protein-DNA recognition**

Several lines of evidence suggest that discrimination of specific DNA sequences by proteins depends in part on the recognition of sequence-dependent differences in DNA structure, such as groove width (Rohs et al., 2010). In the present work, we find that all

preferred binding sites, regardless of Exd-Hox preference, are predicted to have narrow minor grooves at 5'-TGAY-3' (positions 2 to 5), and that the groove tends to stay narrow in the Exd direction, likely due to the presence of short A-tracts in many of the sequences. In all of the existing crystal structures, Arg5 of both Exd and Hox are either bound to or located near to this narrow minor groove region. Arg has been shown to be attracted through electrostatic interactions to narrow minor grooves (Rohs et al. 2009), and all Hox proteins and Exd have an Arg at position 5 of their homeodomains (Figure 3-5 A).

In contrast to this shared feature, minor groove topography varies in the Hox portion of these binding sites. Most notably, anterior Hox proteins select binding sites that have an additional minor groove minimum close to the AY of the Hox half site, 5'-NNAY-3', whereas posterior Hox proteins prefer a wider minor groove in this region. In several cases the binding sites preferred by a particular Exd-Hox complex have similar DNA shapes despite having different sequences, in agreement with the observation that DNA shape is often more conserved than DNA sequence (Parker et al., 2009). That minor groove shape may play an important role in Exd-Hox binding preferences is further underscored by our observation that this parameter was sufficient to accurately partition the preferred binding sites of Hox proteins, irrespective of the primary sequence.

Looking more broadly at the selected binding sites, it is of particular interest that most of the sequence variation contributing to Hox preference is located at positions 6 and 7 (Figure 3-2). Remarkably, the base pair at position 7 makes no protein contacts in any of the known crystal structures, while position 6 makes only a small number of contacts that do not appear to be specific. How is it possible that a single nucleotide

position that makes no contacts can play such an important role in specificity? Replacing a purine at position 7 with a T shifts the location of a TpR step (where R=A or G) in the 3’ direction, which tends to widen the minor groove (Joshi et al, 2007). The TpA or TpG step in most anterior DNA sites (positions 6 and 7) would thus widen the groove in the middle of the binding site, allowing Arg3 and Arg5 to bind to the two minima on either side. In contrast, the TpA step in posterior DNA sites at positions 7 and 8 may block Arg3 from stably inserting into the groove. Notably, the replacement of an A with a G at position 7 forms a CpA step on the opposite strand, which has similar properties to the TpA step thus accounting for the presence of either an A or G at this position in Dfd and Scr. While a more detailed understanding of the role of positions 6 and 7 will benefit from additional crystal structures, the shape analysis presented here nevertheless highlights the general importance of DNA shape for specific DNA binding by Hox proteins.

### 3.3.3 **Constraints on the evolution of Exd-Hox binding preferences**

When the first complex of Hox genes was discovered in *Drosophila*, it was realized that the order of Hox genes along the chromosome was collinear with their corresponding functional domains along the anterior-posterior axis of the adult fly (Lewis, 1978). Collinearity was later extended to Hox expression patterns along the anterior-posterior axis during fly and vertebrate embryogenesis (McGinnis and Krumlauf, 1992). Here, we extend this rule further by showing that differences in DNA-binding specificities of Exd-Hox complexes, as well as the minor groove shape of their preferred binding sites, are also collinear with the characteristics of Hox genes. Collinearity of DNA-binding preferences likely reflects the way in which Hox genes duplicated and

gradually diverged during evolution (Hueber et al., 2010; McGinnis and Krumlauf, 1992).

It is interesting that, when presented with all possible 16-mers, the preferred binding sites for each of the eight Exd-Hox complexes characterized here all share the structure 5'-WRAYNNAY-3'. This binding site matches nearly all of the known *in vivo* binding sites for Exd-Hox or Pbx-Hox complexes (Mann et al., 2009). Thus, it appears that for $Hth^{HM}$-Exd-Hox complexes, alternative modes of binding are not used by these factors. These observations suggest that the biophysical properties of these proteins have constrained the evolution of Exd-Hox-DNA interactions: evolution may have had only a limited potential to modify these interactions. Moreover, the preferred binding sites identified by SELEX-seq are present in bona fide *in vivo* binding sites that have been characterized by more traditional methods: for example, Exd-Scr regulates its target *fkh250* via a blue binding site and Exd-Lab auto-regulate *labial* via a yellow binding site (Ryoo and Mann, 1999; Ryoo et al., 1999). We also found that on a genome-wide level, regions bound *in vivo* by Exd-Ubx are specifically enriched in red and magenta binding sites. Although chromatin structure and interactions with other proteins *in vivo* no doubt also influence Hox binding and activity, these findings suggest that the Exd-Hox binding site signatures identified here will be important for deciphering the sequence determinants that guide the binding, and eventually the function, of these proteins *in vivo*.

## 3.4 Method

### 3.4.1 **DNA shape prediction**

All-atom Monte Carlo (MC) simulations without the protein in presence were

used to predict structural features intrinsic to nucleotide sequence of the DNA targets. The MC simulations were initiated from ideal B-DNA structures of 20-mers that have the 5'-nTGAYNNAYnnn-3' motif in the center of the variable 16-base pair region (excluding reads with more than one motif). The MC simulation protocol was described previously (Joshi et al., 2007). The sampling algorithm is based on collective and internal variables (Rohs et al., 2005b), an analytic chain closure using associated Jacobians (Sklenar et al., 2006), explicit sodium counter ions, and an implicit solvent model described by a distance-dependent sigmoidal dielectric function (Rohs et al., 1999). The Metropolis-Boltzmann criterion was applied based on energy calculations within the framework of the AMBER94 force field (Cornell et al., 1995). Resulting MC trajectories were analyzed with CURVES (Lavery and Sklenar, 1989) in the 5'-TGAYNNAY-3' direction, thereby providing average structural parameters. Independent MC simulations were performed in cases where force field artifacts led to deformations, thus restricting the conformational search to B-DNA. The Mann-Whitney U tests for comparing anterior shape and posterior shape were performed by SciPy's stats.mannwhitneyu module (http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html).

### 3.4.2 **High-throughput DNA shape prediction**

A total of 1,658 trajectories from independent MC simulations were used to build a database for DNA shape predictions. These MC trajectories were analyzed in terms of conformations of all associated tetra- and penta-nucleotides. The data derived from tetramer and pentamer conformations were combined in a hybrid model, which uses only pentamer data if the pentamer occurrence >3, a combination of penta- and tetramer data if the pentamer occurrence ≤3, and only tetramer data if the pentamer occurrence is 0. The

hybrid model is necessary because only 467 of the 512 unique pentamers (91%) occur in our current dataset compared to the almost complete coverage of 135 of the 136 unique tetramers. Each tetra-nucleotide occurs on average 178 times and each penta-nucleotide on average 50 times in the MC data used for the predictions.

Applying this method to the SELEX-seq binding sites, the average minor groove width at the two central nucleotides of tetramers and the central nucleotide of pentamers were used to infer the shape of all sequences that had a relative affinity above 0.1. All reads were aligned based on the 5'-TGAYNNAY-3' motif (excluding reads with more than one motif) and the average minor groove width in each position was calculated. The width values at the six positions of the 5'-AYNNAY-3' core motif were used to calculate a Euclidean distance tree that relates the shapes selected by all Exd-Hox dimers. This dendrogram was generated with the UPGMA method as implemented in the MEGA program (Tamura et al., 2011).

### 3.4.3 **Identification of paralog-specific Hox residues that correlate with specificity**

Sequences for Hox proteins were obtained from NCBI's Protein database (http://www.ncbi.nlm.nih.gov/protein/advanced). The species used were restricted to arthropods and sequences that contain more than 80 amino acids. Except for AbdB, all the Hox sequences must contain a WM motif. In cases where more than one sequence from the same species was present, the longest sequence was used. Protein sequences that contain "predicted", "putative", or "hypothetical" in the headlines of their fasta files were discarded. Homeodomains were identified through the conserved Q50-N51-R52-R53 motif and the YPWM motif was determined by the conserved tryptophan residue across all eight Hox proteins. Partially conserved residues that correlate with DNA-binding

specificity were identified by comparing the multiple sequence alignment (Figure 3-5 A). For each of the four most favored hexamer DNA binding sites (red, blue, green, and magenta motif in Figure 2-3), a subset of Hox proteins was built to include those having relative binding affinity higher than 0.5. Then, for each aligned position, residues that are 100% conserved across Hox proteins in the subset, but have different residue types in the other Hox proteins, were defined as the partially conserved residues that correlate with DNA-binding specificity. Exd-Pb's hexamer preference was not considered during the identification of partially conserved residues. Hydrogen bonds between partially conserved residues and DNA were from contact maps in the papers that describe the crystal structures (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999).

## 3.5 A note on the collaborative works described in this chapter

This chapter is part of a paper published in *Cell* (Volume 147, Issue 6, pp. 1270-1282, December 9, 2011) by Matt Slattery*, Todd Riley*, Peng Liu*, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen Bussemaker, and Richard Mann (*equal contribution). Author contributions: M.S. designed and executed the SELEX experiments and contributed to the analysis of the SELEX data. T.R. designed and executed the analysis of the SELEX-seq data. P.L. carried out and analyzed Monte Carlo simulations. N.A. carried out later rounds of SELEX and EMSA-based validation experiments. P.G.-A. analyzed Ubx ChIP data. I.D. executed high-throughput DNA shape analysis. T.Z. developed high-throughput DNA shape prediction method. R.R. and B.H. supervised and designed DNA shape analyses. H.J.B. supervised

and designed the analysis of the SELEX-seq data. R.S.M. supervised and designed SELEX experiments; contributed to the analysis of the SELEX data and shape analyses.

# Chapter 4. Using Homology Modeling to Infer Minor Groove Recognition Mode of Hox Protein Scr to Its Favored DNA Binding Sites

## 4.1 Introduction

Hox proteins are homeodomain-containing transcription factors that define body plan in both vertebrates and invertebrates (McGinnis and Krumlauf, 1992). Individual Hox gene is expressed at restricted segments along the anterior-posterior axis to specify cellular and tissue identities. Mutations of Hox proteins lead to body malformations, such as the antenna-to-leg transformation in fly (Lewis, 1992) and Hand-Foot-Genital syndrome in human (Goodman and Scambler, 2001). For instance, a missense mutation of a key DNA recognition residue Asn51 to His produces extremely short thumbs and absent halluces (Li et al., 2011). In order to understand the exquisite developmental functions of Hox proteins, it is necessary to study their DNA-binding site preferences and identify the origins for their specificity.

Towards decoding Hox specificity, a variety of experimental approaches have been carried out (Mann et al., 2009). Earlier structural studies demonstrated that Hox proteins use the third α-helix in their homeodomains to recognize DNA major groove (Gehring et al., 1994b). In particular, Asn51 makes direct hydrogen bonds to an adenine and Ile47 forms hydrophobic contact to a thymine. Measurements from high-throughput methods, such as protein binding microarray (Berger et al., 2008) and bacteria one-hybrid (Noyes et al., 2008a), confirmed findings from the structural studies. The homeodomains

of Hox proteins from fly and mouse were uncovered to favor a TAAT motif (Berger et al., 2008; Noyes et al., 2008a), where $A_3$ and $T_4$ are the ones specifically recognized by Asn51 and Ile47. Despite these advances on the recognition of major groove, a consensus TAAT motif cannot fully explain the diverse regulatory functions of Hox proteins in controlling body development (Affolter et al., 2008; Mann et al., 2009).

The recognition of DNA minor groove was demonstrated as another origin for Hox specificity from the comparison of two crystal structures (Joshi et al., 2007). Both structures contain the same Hox protein, Sex comb reduced (Scr), and the same cofactor, Extradenticle (Exd). The difference is in that one binds to Scr's *in vivo* specific site *fkh250*, while the other binds to Scr's non-specific site $fkh250^{con}$. The extra narrow region in *fkh250*'s minor groove was found to induce enhanced negative electrostatic potential and thereby attracts Scr's Arg3 to bind, resulting in Scr's specificity to *fkh250*. Further studies uncovered that this type of minor groove shape preference is a widely used mechanism for proteins to recognize their specific binding sites (Rohs et al., 2010; Rohs et al., 2009b), especially for the nucleosome core particles (West et al., 2010). Recently, a high-throughput SELEX-seq experiment revealed that minor groove shape is a structural determinant for the specificities of all eight *Drosophila* Hox proteins (Slattery et al., 2011b). The four anterior Hox prefer a type of minor groove shape similar to *fkh250*, whereas the four posterior Hox favor a type of shape similar to $fkh250^{con}$. This difference in shape preference raises the question of how particular minor groove is recognized by Hox proteins.

One way to answer this question is through structure-based homology modeling, which has been broadly used to study protein-DNA interactions. Siggers and Honig

developed an all-atom homology modeling method to predict the binding specificity of $C_2H_2$ zinc finger proteins. They found that the prediction accuracy largely depends on the similarity of protein-DNA interface geometry between template and target (Siggers and Honig, 2007). The structure-based modeling software, Rosetta, was developed to redesign DNA-binding specificities of endonuclease I-MsoI and I-AniI (Ashworth et al., 2006; Thyme et al., 2009). A threading-based method, DBD-Threader, was demonstrated to achieve considerably high sensitivity and precision in predicting DNA-binding domains and associated DNA-binding residues (Gao and Skolnick, 2009). The structure-based protein design algorithm FoldX, was shown to predict 88% of Pax6's mutations involved in human disease and reproduce experimentally determined motifs (Alibes et al., 2010). Morozov and Siggia used the residue-nucleotide contacts from homology models to refine binding sites for 67 transcription factors and correctly assign transcription factor families to available binding sites (Morozov and Siggia, 2007). However, in spite of all these successful approaches to decipher protein-DNA binding specificity, none of them has investigated the specific recognition of protein to DNA minor groove shape.

Here, we develop a structure-based homology modeling method, iPRED (interface Prediction for REcognition of Dna), to infer the minor groove recognition mode for Hox protein Scr. This method is based on the all-atom modeling approach Siggers and Honig used on the study of $C_2H_2$ zinc fingers (Siggers and Honig, 2007). We first briefly introduce iPRED's protocol of protein-DNA docking and interface optimization. Then, we evaluate iPRED's energy function by checking whether it can capture sequence-dependent electrostatic feature in DNA minor groove. We also examine

iPRED's side-chain prediction method on residues that recognize minor groove shape. Next, iPRED is validated on the crystal structures of Exd-Scr-*fkh250* and Exd-Scr-*fkh250*$^{con}$ to see whether it can reproduce the native minor groove recognition modes. Last, we apply iPRED to build homology models for Scr's preferred sites obtained from SELEX-seq experiment. The structural basis of how Scr recognizes the minor groove shape of these sites is inferred from homology models.

## 4.2 Results

### 4.2.1 **Overview of iPRED**

Our homology modeling method for protein-DNA docking and interface optimization consists of five steps (Figure 4-1, see Materials and Methods for more details). First, for a given SELEX-seq sequence, its structure is predicted by Monte Carlo (MC) simulation (Rohs et al., 2005b; Sklenar et al., 2006). The predicted DNA is superposed onto a template structure of Scr-Exd-DNA complex to produce a hybrid model where Scr and Exd are from the template and DNA is from MC prediction. Then, a moderate energy minimization (root mean square gradient of 1.0 kcal/mol/Å) is performed on the hybrid model to remove steric clashes at protein-DNA interface. Next, backbone of the first six residues on Scr's N-terminal arm is perturbed locally to sample 1,000 conformations. Side-chain conformations of the six residues are predicted for each sampled backbone, followed by a rigorous energy minimization (root mean square gradient of 0.01 kcal/mol/Å). Finally, structural model with the lowest conformational energy is selected as the predicted structure, from which the minor groove recognition mode is inferred.

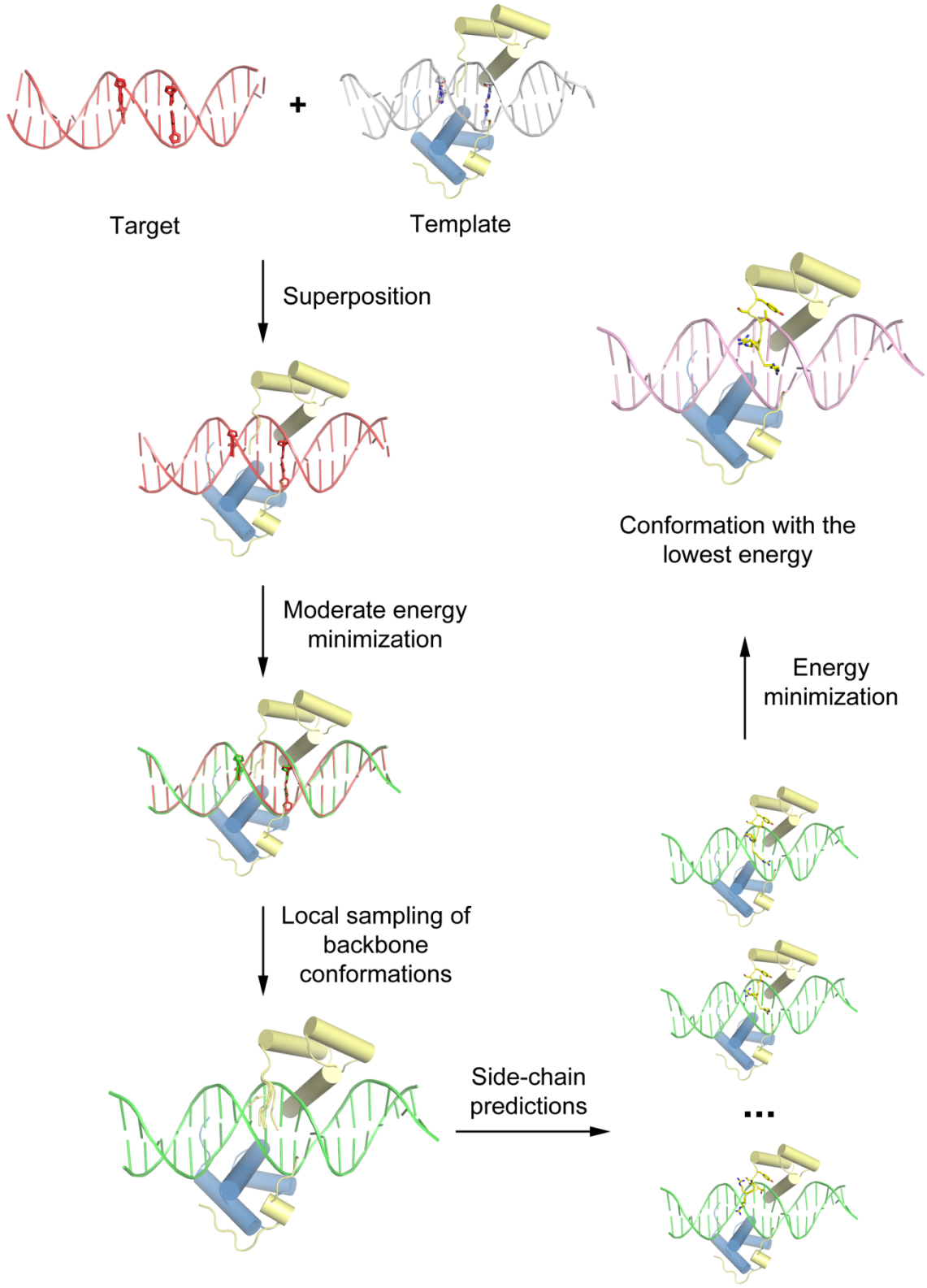


**Figure 4-1. Overview of the iPRED method.**

The flowchart illustrates the procedure of protein-DNA docking and interface optimization in iPRED. Protein colored in yellow is represented as cylinders and loops. DNA is denoted as double-strand ladder. Protein side-chains are shown in sticks.

Before applying iPRED to infer the minor groove recognition mode of Scr to its favored SELEX-seq sequences, we want to see whether it can reproduce the minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$ as observed from crystal structures. Therefore, we carried out three validation tests to ask: i) whether iPRED's energy function can capture sequence-dependent electrostatic potential in minor groove; ii) to what accuracy iPRED's side-chain prediction can performed for residues recognizing narrow minor groove; and iii) whether iPRED can reproduce the minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$.

### 4.2.2 Sequence-dependent electrostatic feature of DNA minor groove is captured by iPRED's energy function

Since the sequence-dependent electrostatic potential in DNA minor groove determines Scr's specificity, we asked the question whether iPRED's energy function can capture such feature. In iPRED, electrostatic potential is calculated by a combination of a softened Coulombic term, a sigmoidal distance-dependent dielectric function describing solvent screening, and charge-reduced phosphate groups accounting for salt effects (Siggers and Honig, 2007). We first tested iPRED's energy function on the two Scr-binding sites: *fkh250* and *fkh250*$^{con}$. Electrostatic potentials in DNA minor groove were calculated in the same way as before (Rohs et al., 2009b). Calculated results were compared with the ones from solving Poisson-Boltzmann equation (PB), which has been

demonstrated to accurately detect the electrostatic feature in DNA minor groove (Jayaram et al., 1989; Rohs et al., 2009b). For electrostatic potentials in *fkh250*'s minor groove, the local minima at $A_4T_5T_6$ and $A_7A_8T_9$ revealed by PB were also observed from the calculations by iPRED's energy function (Figure 4-2 A). The agreement holds true for the local minimum at $A_4T_5T_6$ of $fkh250^{con}$ as well (Figure 4-2 B). Although a slight minimum was reported at $A_7A_8T_9$ in $fkh250^{con}$ by iPRED's energy function (Figure 4-2 B), the lowest value is -10.16 kT/e, which is still 1 kT/e higher than the lowest value for the corresponding region in *fkh250* (Figure 4-2 A), indicating that iPRED's energy function can detect the difference of electrostatic potentials between *fkh250* and $fkh250^{con}$.

Next, we extended the comparison to all five Hox-cofactor binding sites (Figure 4-2). All comparisons have correlation coefficients higher than 0.6 and p-values lower than 0.05 (Table 4-1), indicating a good agreement between iPRED's energy function and PB. The moderate correlation on the Exd-Ubx binding site is most likely due to the relatively short DNA sequence, which is only 13 base-pairs, in the Exd-Ubx-DNA structures, because the other four Hox-binding sites all have a 19 base-pairs DNA. Since the probes for measuring electrostatic potential are usually close to the hydrogen atoms of guanine's NH2 group in minor groove, iPRED's energy function appears to overestimate the electrostatic potential at GC base-pair (Figure 4-2). Nevertheless, most of the comparisons demonstrate that electrostatic potentials calculated by iPRED's energy function have strong correlation with PB calculations, suggesting that the sequence-dependent electrostatic feature in DNA minor groove can be captured by iPRED's energy function.

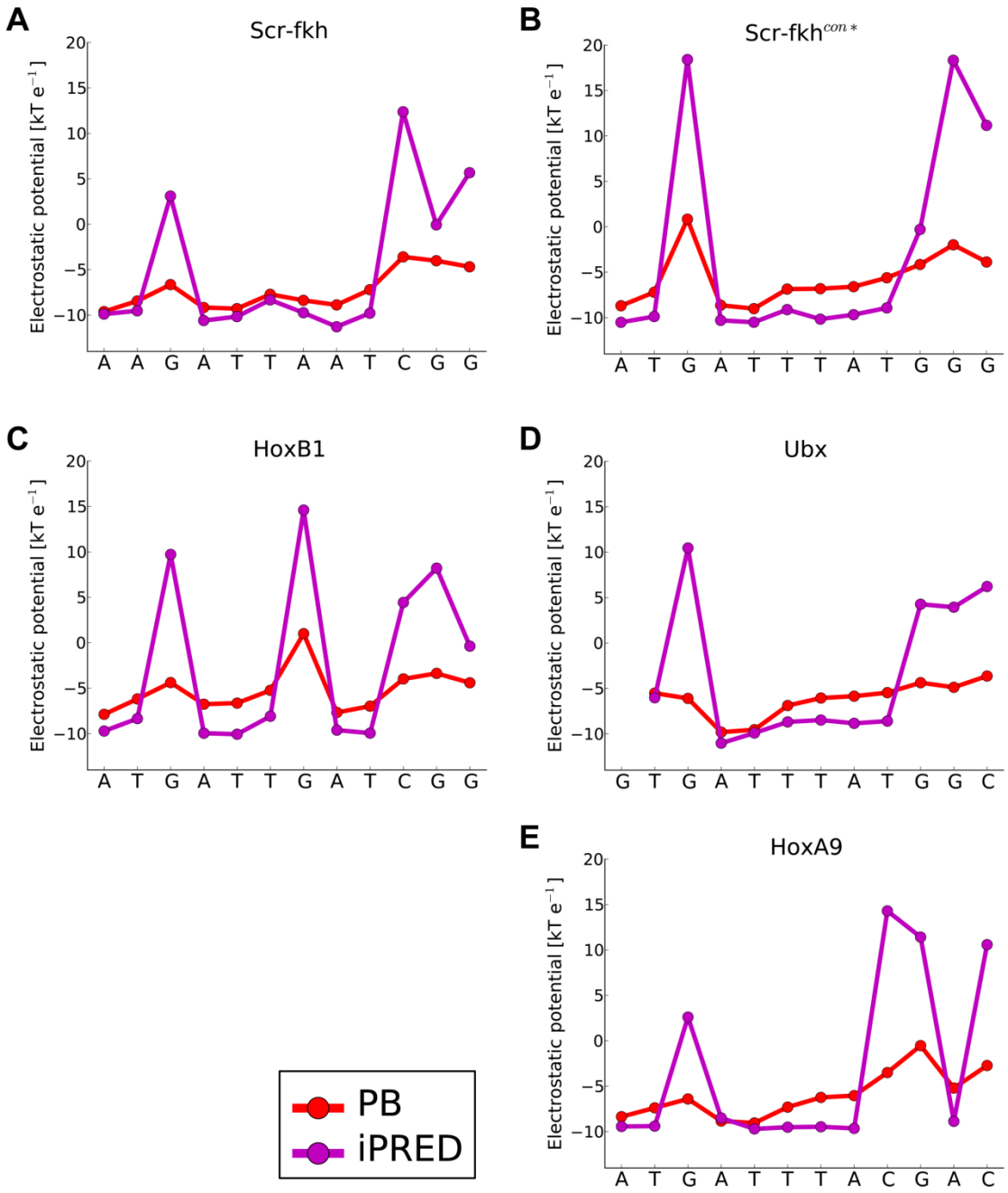


**Figure 4-2. Sequence-dependent electrostatic feature in DNA minor groove is captured by iPRED's energy function.**

(A-E) Comparison of electrostatic potential calculated by iPRED's energy function (magenta) and Poisson-Boltzmann equation (red) for DNA-binding sites from five Hox-cofactor-DNA complexes.

The identity of each Hox-cofactor-DNA structure is denoted on top of each plot (See Table 4-1 for abbreviations).

**Table 4-1. Electrostatic potentials calculated by iPRED's energy function have strong correlations with the ones solved by Poisson-Boltzmann equation.**

| Structure[a] | Correlation coefficient | p-value |
|---|---|---|
| HoxB1 | 0.90 | $8.4 \times 10^{-5}$ |
| Scr-fkh | 0.89 | $9.3 \times 10^{-5}$ |
| Scr-fkh$^{con}$ | 0.92 | $2.7 \times 10^{-5}$ |
| Ubx | 0.62 | $4.1 \times 10^{-2}$ |
| HoxA9 | 0.83 | $9.2 \times 10^{-4}$ |

[a] Abbreviations were used in Table 4-1, 4-2 and Figure 4-2, 4-3 to represent the five Hox-cofactor-DNA structures. HoxB1: HoxB1-Pbx-DNA (Piper et al., 1999); Scr-fkh: re-refined Scr-Exd-*fkh250* (Joshi et al., 2007); Scr-fkh$^{con}$: re-refined Scr-Exd-*fkh250*$^{con}$ (Joshi et al., 2007); Ubx: Ubx-Exd-DNA (Passner et al., 1999); HoxA9: HoxA9-Pbx-DNA (LaRonde-LeBlanc and Wolberger, 2003).

### 4.2.3 Side-chains for minor groove recognition are predicted as accurate as those for major groove recognition

Prediction of side-chain conformations is an essential component in iPRED. To validate whether iPRED can correctly repack side-chains that recognize minor groove shape, we carried out a test on the His-12 and Arg3 from Exd-Scr-*fkh250* structure, and the Arg5 from all five Hox-cofactor-DNA structures. The predicted conformations of His-12 and Arg3 were close to the native structure (Figure 4-3 A). Both have RMSDs less than 0.6 Å and correctly predicted $\chi_1/\chi_{1+2}$ (Table 4-2). The predicted Arg5 has a relatively higher RMSD and lower $\chi_1/\chi_{1+2}$ (Table 4-2), yet it is still in contact with DNA minor groove as the native conformation illustrated in the Exd-Scr-*fkh250* structure (Figure 4-3 A).

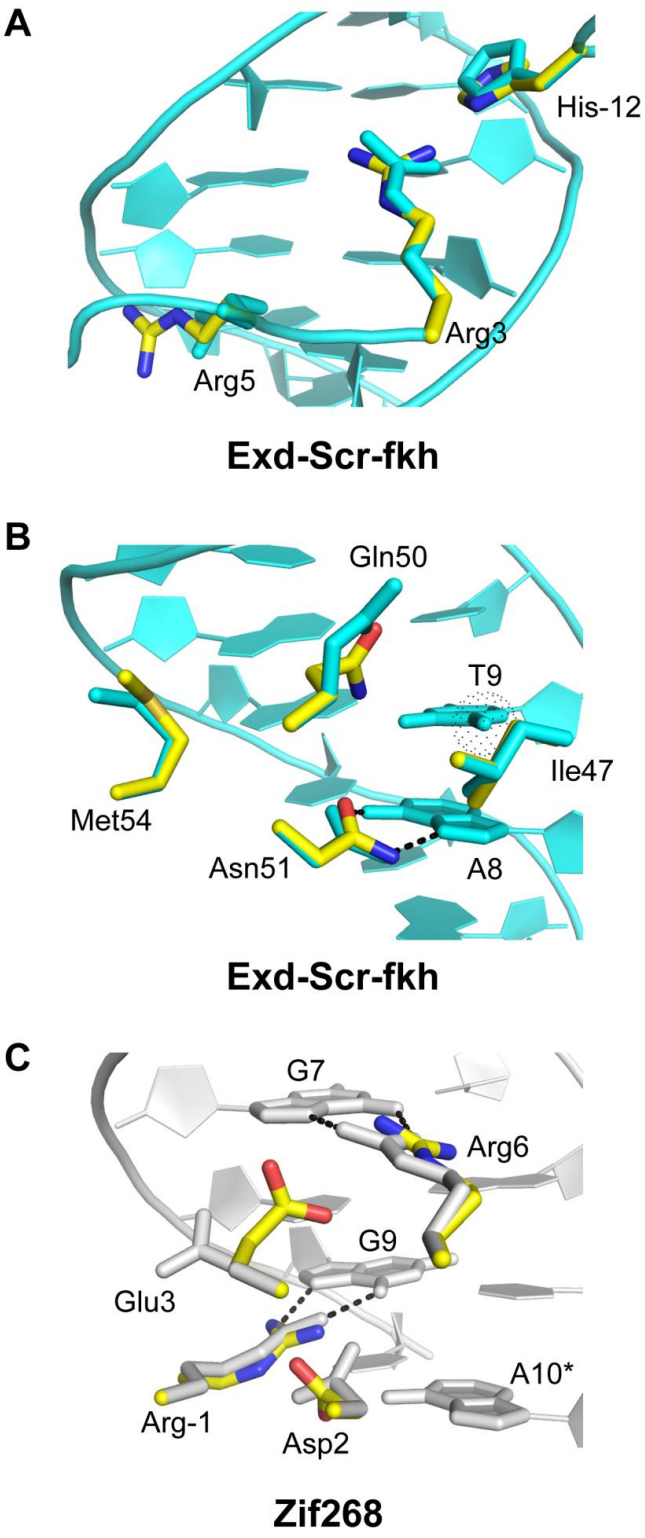


**Figure 4-3. Side-chains for minor groove recognition are modeled at the same accuracy level as those for major groove recognition.**

(A-C) Comparison of native conformations (cyan in A and B, gray in C) with repacked side-chains (sticks in yellow, blue, and red) for key residues that recognize *fkh250*'s minor groove (A), *fkh250*'s major groove (B), and Zif268-binding site's major groove (C). Direct hydrogen bonds between native side-chains and DNA bases are represtend in dash lines. Hydrophobic interaction is shown in dotted sphere.

**Table 4-2. Side-chains for minor groove recognition are modeled on the same accuracy level as those for major groove recognition.**

| | Structure | Residue | RMSD [Å] | $\chi_1/\chi_{1+2}$ [%/%] |
|---|---|---|---|---|
| **Hox minor groove** | Scr-fkh | His-12 | 0.59 | 100/100 |
| | Scr-fkh | Arg3 | 0.43 | 100/100 |
| | All 5 Hox | Arg5 | 1.53 | 60/40 |
| **Hox major groove** | All 5 Hox except HoxA9[a] | Ile47 | 0.29 | 100/100 |
| | All 5 Hox | Gln50 | 2.32 | 60/40 |
| | All 5 Hox | Asn51 | 0.61 | 100/60 |
| | All 5 Hox | Met54 | 2.42 | 80/40 |
| **Zif268 major groove** | ZF 1,2,3[b] | Arg-1 | 0.36 | 100/100 |
| | ZF 1,2,3 | Asp2 | 0.69 | 100/67 |
| | ZF 1,2,3 | Glu3/His3 | 1.68 | 100/67 |
| | ZF 1,2,3 | Arg6/Thr6 | 0.71 | 100/100 |

[a] The prediction results on HoxA9's Ile47 are excluded, because its side-chain has alternative conformations in the PDB structure (LaRonde-LeBlanc and Wolberger, 2003). [b] "ZF 1, 2, 3" denotes the three zinc finger domains in Zif268 (Elrod-Erickson et al., 1996).

To compare the prediction accuracy, we repacked key DNA-contacting residues in major groove from five Hox-cofactor-DNA structures and a $C_2H_2$ zinc finger Zif268-DNA structure. The predictions of major groove contacting residues were found to have a similar accuracy level as predictions of minor groove contacting residues. Hox's Ile47 and Asn51 along with Zif268's Arg-1, Asp2, and Arg6/Thr6 recognize DNA bases through direct hydrogen bonds or hydrophobic contacts (Figure 4-3 B and C). Their

predicted side-chain conformations have RMSDs around or lower than 0.7 Å (Table 4-2), which is on the same accuracy level as the predictions for Scr's His-12 and Arg3. Hox's Gln50 and Met54 as well as Zif268's Glu3/His3 make water-mediated hydrogen bonds or non-specific contacts to DNA bases in major groove (Figure 4-3 B and C). Predictions on their conformations show RMSDs ranging from 1.7 Å to 2.4 Å (Table 4-2), which are similar or even higher than predictions on Hox's Arg5. The comparisons of side-chain predictions show that side-chains for recognizing minor groove shape are predicted at a similar accuracy level as side-chains that recognize major groove.

### 4.2.4 The native minor groove recognition modes of Scr to *fkh250* and *fkh250$^{con}$* are reproduced by iPRED

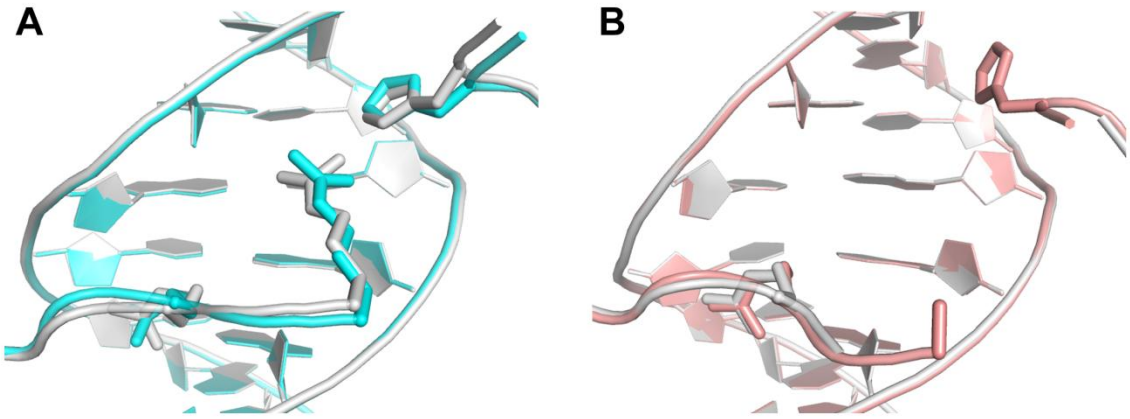


**Figure 4-4. Minor groove recognition modes in re-refined Scr-Exd-*fkh250* and Scr-Exd-*fkh250$^{con}$* structures are consistent with previously published structures.**

(A and B) Comparison of re-refined Scr-Exd-*fkh250* (cyan) and Scr-Exd-*fkh250$^{con}$* (light pink) with previously published structures (Joshi et al., 2007) (gray). His-12, Arg3, and Arg5 are shown in sticks.

To check whether iPRED can reproduce the native minor groove recognition mode, we tested it on the Scr-Exd-*fkh250* and Scr-Exd-*fkh250$^{con}$* structures. Both

structures have been re-refined recently based on their electron density maps. The re-refined Scr-Exd-*fkh250* structure remains almost the same as the previous one, in which His-12, Arg3, and Arg5 recognize the narrow minor groove (Figure 4-4 A). In the re-refined Scr-Exd-*fkh250*$^{con}$ structure, on the other hand, His-12 becomes visible and contacts minor groove (Figure 4-4 B). The side-chain of Arg3, except the Cβ atom, is still disordered (Figure 4-4 B), confirming previous finding that Arg3 is not engaged in minor groove recognition.

To avoid bias in the template structure, we averaged the backbone conformations of Scr from Scr-Exd-*fkh250* and Scr-Exd-*fkh250*$^{con}$ structures and thus generated a template, where the N-terminal arm locates between the two native structures (Figure 4-5 A). Starting from this template, we tested iPRED by inferring the minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$. Both Arg3 and Arg5 bind to the minor groove of *fkh250* (Figure 4-5 B). To quantitatively characterize such binding, we measured Arg's "CZ contact distance", defined as the minimum distance between Arg's Cζ atom and DNA base atoms in minor groove. In the native structure of Scr-Exd-*fkh250*, Arg3 and Arg5 bind to minor groove with CZ contact distances in 5.5 Å and 3.4 Å, respectively (Table 4-3). Same range of CZ contact distances were observed from our models for both residues (Table 4-3), indicating that iPRED correctly reproduced Scr's minor groove recognition mode to *fkh250*. For *fkh250*$^{con}$, only Arg5 binds to the minor groove in our model (Figure 4-5 C) with a CZ contact distance close to the native (Table 4-3). Arg3, in contrast, is away from minor groove (Figure 4-5 C) with a CZ contact distance more than 10 Å (Table 4-3). The two validation tests show that iPRED is able to predict the minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$.

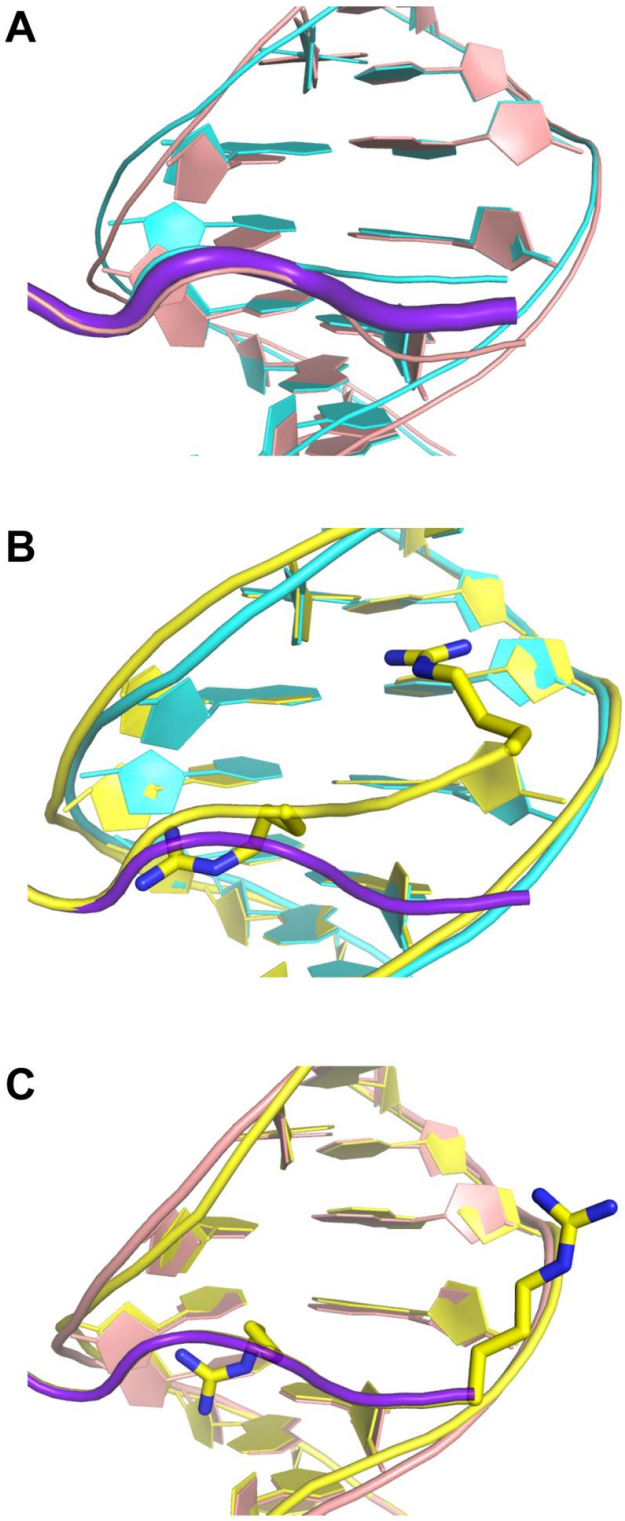


**Figure 4-5. Minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$ are reproduced by iPRED.**

(A) A template structure of Scr (magenta) is generated by averaging backbone conformations from the re-refined Scr-Exd-*fkh250* (cyan) and Scr-Exd-*fkh250*$^{con}$ (light pink) structures. (B and C) Homology models (yellow) for inferring minor groove recognition modes of Scr to *fkh250* (B) and *fkh250*$^{con}$ (C) were built based on the template structures (magenta and cyan in B, magenta and light pink in C). Side-chains of Arg3 and Arg5 are represented as sticks in yellow and blue.

**Table 4-3. The minor groove recognition modes of Scr to *fkh250* and *fkh250*$^{con}$ are reproduced by iPRED.**

| Residue | Scr-*fkh250* [Å] | | Scr-*fkh250*$^{con}$ [Å] | |
|---|---|---|---|---|
| | X-ray | Model | X-ray | Model |
| Arg3 | 5.5 | 3.6 | N/A[a] | 10.3 |
| Arg5 | 3.4 | 4.1 | 2.8 | 3.1 |

[a] Arg3's side-chain, except Cβ atom, is disordered in the re-refined Scr-Exd-*fkh250*$^{con}$ structure. Thus, no CZ contact distance could be measured.

### 4.2.5 Both Arg3 and Arg5 are used to recognize the minor grooves of Scr's favored sites that have ATTAAT, ATTGAT, or ATAAAT core motifs

SELEX-seq experiment had shown that Exd-Scr prefer to bind DNA sites that have ATTAAT, ATTGAT, ATAAAT, or ATTTAT core motifs, among which ATTAAT is the most favored one (Slattery et al., 2011b). Hence, we started with the highest affinity SELEX-seq sequence that has an ATTAAT core motif, and applied iPRED to study how Scr recognizes the minor groove shape of this sequence. The model built by iPRED shows that both Arg3 and Arg3 bind to the minor groove of the target DNA site (Figure 4-6 A). Extending the study to the top ten high-affinity sites with an ATTAAT core motif, we found that nine out of ten sites have Arg3 in CZ contact distance around 4 Å (Figure 4-6 E), and all ten sites have Arg5 in CZ contact distance less than 4 Å (Figure

4-6 F), suggesting that both Arg3 and Arg5 are used to recognize the minor grooves of Scr-favored ATTAAT sequences.

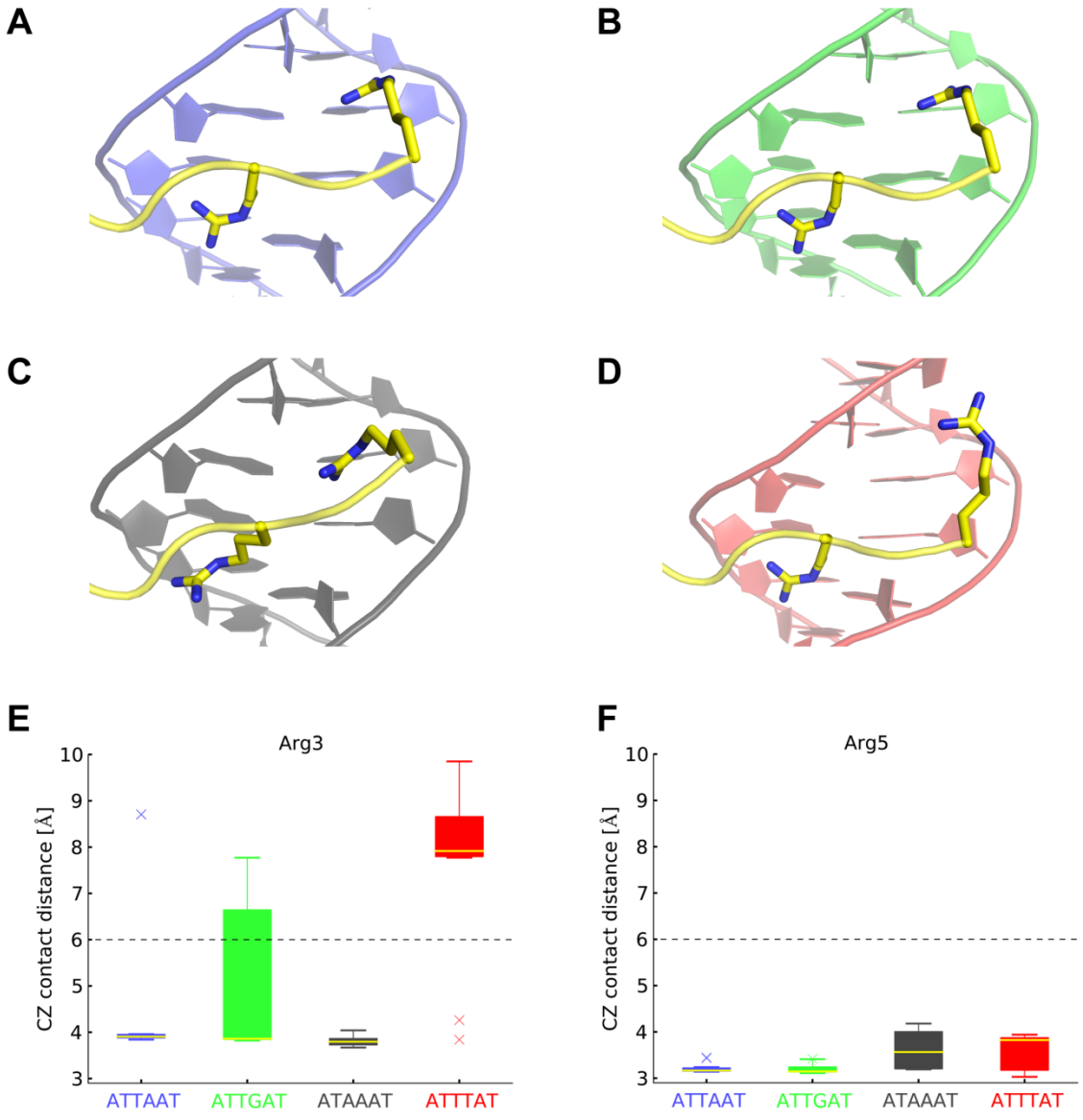


**Figure 4-6. Scr's minor groove recognition modes to high-affinity sequences are inferred from homology models.**

(A-D) Homology models of Scr (yellow) binding to its most favored SELEX-seq sequences having core motifs: ATTAAT (A), ATTGAT (B), ATAAAT (C), and ATTTAT (D). Side-chains of Arg3 and Arg5 are represented as sticks in yellow and blue. (E and F) Box plots showing CZ contact distances

for Arg3 (E) and Arg5 (F) in homology models of Scr binding with the ten most favored SELEX-seq sequences in four types of core motifs. The mean values are highlighted in yellow lines and outliers are represented by "x".

Next, we investigated Scr's minor groove recognition mode to its favored sequences with ATTGAT or ATAAAT motifs. Since both motifs were uncovered before to have an "anterior shape" similar to *fkh250* (Slattery et al., 2011b), it is most likely that both Arg3 and Arg5 are enrolled in minor groove shape recognition. As expected, for the highest affinity sequences that either have an ATTGAT or ATAAAT core motif, both Arg3 and Arg5 bind to their minor grooves (Figure 4-6 B and C). A similar finding was observed by further modeling on the top ten high-affinity sites. Arg3 and Arg5 both have CZ contact distances around 4 Å to the top ten ATAAAT sites (Figure 4-6 E and F). For all the top ten ATTGAT sites, Arg5 has CZ contact distances around 3 Å (Figure 4-6 F). Arg3, on the other hand, does not show a uniformly short CZ contact distance (Figure 4-6 E). In three ATTGAT sites, Arg3 has CZ contact distances more than 7 Å, possibly because the steric effect of guanine's $NH_2$ group in DNA minor groove reduces the number of favorite conformations that Arg3 could adopt. Since all these three sites have relative affinities lower than six out of the ten ATTGAT sites, we can conclude that both Arg3 and Arg5 are engaged in the minor groove recognition to high-affinity ATTGAT and ATAAAT sites.

### 4.2.6 **Only Arg5 is involved in the minor groove recognition to Scr's high-affinity sites with an ATTTAT core motif**

Scr's high-affinity ATTTAT sites have been demonstrated to have a "posterior shape" similar to *fkh250*$^{con}$ (Slattery et al., 2011b). We hypothesized that Scr would recognize these SELEX-seq sites in the same way as it recognizes *fkh250*$^{con}$, where only Arg5 is used to contact the minor groove. To test this hypothesis, we built homology models for the highest affinity ATTTAT site. As expected, Arg5 binds to the minor groove, whereas Arg3 is positioned away from minor groove (Figure 4-6 D). Modeling on the top ten high-affinity ATTTAT sites confirmed Arg5's role. It binds to minor grooves of all the ten sites with CZ contact distances less than 4 Å (Figure 4-6 F). By contrast, Arg3 have CZ contact distances higher than 7.5 Å for eight out of the ten sites (Figure 4-6 E), suggesting that it is not used specifically to recognize the minor groove. Together, only Arg5 is engaged in the recognition of Scr-favored ATTTAT sites.

## 4.3 Discussion

We have developed a new homology modeling method, iPRED, to infer the minor groove recognition mode for Hox protein Scr. The energy function in iPRED is capable to capture the sequence-dependent electrostatic feature of DNA minor groove. Validation tests reproduce the native recognition modes observed in Scr-Exd-*fkh250* and Scr-Exd-*fkh250*$^{con}$ structures. Further studies on Scr's high-affinity SELEX-seq sequences uncovered that Scr's Arg3 and Arg5 recognize the minor groove shape of sites with core motifs ATTAAT, ATTGAT, and ATAAAT. By contrast, only Arg5 is employed to recognize the minor groove of sites that have an ATTTAT core motif. Our work provides an effective way to study the minor groove recognition of Hox proteins to their favored DNA sites and brings more structural explanations for Hox specificity.

### 4.3.1 **A novel homology modeling method for protein-DNA interaction**

Despite many successful homology modeling works have been carried out over the years to study protein-DNA interaction (Alibes et al., 2010; Ashworth et al., 2006; Gao and Skolnick, 2009; Morozov et al., 2005; Morozov and Siggia, 2007; Siggers and Honig, 2007; Thyme et al., 2009), few of them focused on the recognition of DNA minor groove shape. DNA minor groove width has been demonstrated as a structural determinant for Scr and all the other *Drosophila* Hox proteins (Joshi et al., 2007; Slattery et al., 2011b). The recognition of narrow minor groove was shown as a widely used mechanism for many protein superfamilies (Rohs et al., 2009b). The major difficulties to incorporate DNA minor groove shape into homology models are the extensive sampling of DNA conformations and the representation of sequence-dependent electrostatic potential. To overcome these two obstacles, our method takes advantage of a MC simulation method (Rohs et al., 2005b; Sklenar et al., 2006) and an effective way to calculate electrostatic potential (Siggers and Honig, 2007). All initial DNA structures in our models were taken from MC simulations, which have been demonstrated to sample DNA intrinsic conformations effectively for both Scr's specific and nonspecific sites as well as other DNA sequences (Joshi et al., 2007; Rohs et al., 2005a; Rohs et al., 2005b; Rohs et al., 2009a). Electrostatic potentials computed by our energy function have strong correlation with the ones from PB, indicating that our method is capable to capture the sequence-dependent electrostatic feature. Both advantages facilitate the incorporation of minor groove shape into homology modeling.

Side-chain prediction is another essential component for modeling minor groove recognition. Prediction of side-chain conformations has been studied for decades (Krivov

et al., 2009; Ponder and Richards, 1987). Benchmark tests have been carried out on monomeric proteins (Jacobson et al., 2002; Xiang and Honig, 2001; Xiang et al., 2007), protein-protein interfaces (Wang et al., 2005), and recovery of native amino acid identity at protein-DNA interface (Havranek et al., 2004). However, to what extend can we predict side-chain conformations at protein-DNA interface is still poorly understood. In our predictions on the key residues for DNA recognition, we found that specific DNA-contacting residues can be predicted at RMSD lower than 1.0 Å and non-specific contacting residues have RMSD around 2.4 Å. In comparison, prediction accuracy for monomeric proteins is: core residues ~0.9 Å; partially buried residues ~1.9 Å; surface residues ~2.4 Å (Xiang et al., 2007). Thus, specific DNA-contacting residues can be modeled on the same accuracy level as core residues in monomeric proteins. Predictions for non-specific DNA-contact residues are at similar accuracy as partially buried residues. This accuracy level is consistent with the solvent exposure of protein-DNA interface residues, because those fully exposed to solvent are not close to the bottom of DNA groove and therefore are not able to make specific contact to DNA.

Our DNA-docking and interface-optimization method provides a flexible platform for a variety of homology modeling applications. The initial free DNA structure can be taken from MC simulations or other types of prediction methods, such as coarse-grained simulation (Chen et al., 2010) or high-throughput predictions (Slattery et al., 2011b). Our method can be applied not only for inferring minor groove recognition, but also for studying the major groove recognition. Moreover, it can be employed to model protein-RNA and protein-protein interfaces as well.

### 4.3.2 **The structural basis of minor groove recognition for other Hox proteins**

The minor groove recognition mode of Hox protein Scr to its favored SELEX-seq sites has been inferred in this work. The next question would be what the structural basis of minor groove recognition is for the other Hox proteins. Ultrabithorax (Ubx), for example, is a posterior Hox protein that was identified to prefer sequences with core motifs ATTTAT and ATTTAC (Slattery et al., 2011b). Current available structure only reveals the recognition mechanism in major groove, not minor groove. Both Arg3 and Gly4 on Ubx's N-terminal arm are disordered and their side-chain conformations are not determined (Passner et al., 1999). To understand Ubx's minor groove recognition mode, we can apply our homology modeling method in a similar manner as described in this work. Ubx's Arg3 and Gly4 together with Arg5 form an RGR motif, which has been found in the minor groove recognition from several nuclear receptor structures (Huth et al., 1997; Meinke and Sigler, 1999). It is feasible to transplant the RGR conformation to Ubx and build homology models as a way to study minor groove recognition. Similar transplantations combined with homology modeling can be employed for other Hox proteins as well and the structural basis can be inferred to understand their specificity.

## 4.4 Methods

### 4.4.1 **Protein-DNA docking and interface optimization**

For any given SELEX-seq DNA sequence, its structure is predicted by a MC simulation method (Rohs et al., 2005b; Sklenar et al., 2006). The predicted DNA structure is superposed onto template DNA by aligning base-pairs at position 4 and 8 (in the 12-mer numbering described previously (Slattery et al., 2011b)), because these two positions are conserved across all Scr-favored sites. To avoid any bias, a template

structure of Scr was built by averaging the backbone conformations from Scr-Exd-*fkh250* and Scr-Exd-*fkh250*$^{con}$ structures. This template structure was used to validate Scr's recognition mode to *fkh250* and *fkh250*$^{con}$. It was also employed to build homology models for SELEX-seq sequences with ATTAAT, ATTGAT, ATAAAT, or ATTTAT core motifs. His-12 is set to be protonated at Nε atom during the validations since both Scr-Exd-*fkh250* and Scr-Exd-*fkh250*$^{con}$ structures were solved under PH 8.7 (Joshi et al., 2007). Both types of protonation at Nε and Nδ atoms were used when building homology models for SELEX-seq sequences.

After the superposition, a moderate energy minimization is performed by Tinker (http://dasher.wustl.edu/ffe/) with a root-mean-square (rms) gradient of 1.0 kcal/mol/Å. During minimization, only backbone atoms of Scr and Exd were fixed, while all the other atoms are allowed to move. A wriggling algorithm (Cahill et al., 2003) was used to sample 1,000 backbone conformations for Scr's residues Arg3 to Tyr8. For each sampled backbone conformation, the side-chains of Arg3 to Tyr8 are repacked. The base-pair conformations of DNA core motif AYNNAY were repacked as well. Detailed side-chain prediction algorithm is described below. A rigorous energy minimization is performed on each model by Tinker with a default rms gradient of 0.01 kcal/mol/Å to optimize residues of Arg3 to Tyr8 and AYNNAY motif. An energy function described before (Siggers and Honig, 2007) is used to evaluate each model. The one with the lowest conformational energy is taken as the final predicted structural model.

### 4.4.2 **Side-chain prediction: algorithm and evaluation of accuracy**

The energy function, rotamer library, and sampling procedure are the same as described before (Siggers and Honig, 2007). A screening was performed in the beginning

of prediction to remove rotamers that has steric clashes with fixed atoms. A steric clash occurs when d < F*(Ri+Rj), where d is the distance between the two atoms; F is the overlapping factor, which is set to 0.75; Ri and Rj are the van der Waals radii for atom i and j, respectively. For each side-chain, up to 30 lowest energy rotamers are kept and 100 initial random conformations are used for the iterative sampling of side-chain conformations. For each nucleotide, rotamers were generated by sampling ±10° of χ angles from the input structure at an interval of 0.5 °.

The accuracy of side-chain prediction was assessed in terms of root-mean-square deviation (RMSD) and $\chi_1/\chi_{1+2}$ by comparing predicted conformation to the native. Cβ atom and hydrogen atoms were not included when calculating RMSD. $\chi_1/\chi_{1+2}$ represent prediction accuracy for side-chain torsion angles. A torsion angle is defined as correctly predicted if it is within 40° to the native. $\chi_1$ is the percentage of side-chains that have the first torsion angle correctly predicted. $\chi_{1+2}$ is the percentage of both first and second torsion angle correctly predicted. Symmetry of side-chain torsions for Asp, Arg, Glu, Phe and Tyr were considered and the lower RMSD and the higher $\chi_{1+2}$ were taken.

### 4.4.3 **Calculation of electrostatic potential, definition of Arg's CZ contact distance, and data set of X-ray structures**

The probes for measuring electrostatic potential and the way to solve PB equations are the same as before (Joshi et al., 2007). Softened Coulombic interactions and the sigmoidal distance-dependent electrostatic model was described previously (Siggers and Honig, 2007). DNA sites are the 12-mer base-pair consisting of the 6-mer core motif, cofactor flank, and Hox flank defined previously (Slattery et al., 2011b). Correlation coefficients and p-values were calculated by the "peasonr" function in SciPy

(http://www.scipy.org/doc/api_docs/SciPy.stats.stats.html). Arg's CZ contact distance is defined as the shortest distance between Arg's Cζ atom to DNA base heavy atoms in the minor groove. The five Hox-cofactor-DNA structures and the $C_2H_2$ zinc finger structure used in this work are: Pbx-HoxB1-DNA (PDB ID: 1b72 (Piper et al., 1999)), re-refined structures based on previously published Exd-Scr-*fkh250* (PDB ID: 2r5z) and Exd-Scr-*fkh250*$^{con}$ (PDB ID: 2r5y) structures (Joshi et al., 2007), Exd-Ubx-DNA (PDB ID: 1b8i (Passner et al., 1999)), Pbx-HoxA9-DNA (PDB ID: 1puf (LaRonde-LeBlanc and Wolberger, 2003)), and Zif268 (PDB ID: 1aay (Elrod-Erickson et al., 1996)).

# Chapter 5. Towards Automatically Identifying Key Interactions in Protein-DNA Structures: An Annotation Module in MarkUs Server

## 5.1 Introduction

The recognition of Hox proteins to their DNA sites have been studies in previous chapters. Structural analysis on all five available Hox-cofactor-DNA complexes has revealed conserved features in Hox-DNA interactions (Slattery et al., 2011b). All five Hox proteins use their Asn51 to make bidentate hydrogen bonds to an adenine in the major groove, and all five Hox proteins, except HoxB1, contact DNA's narrow minor groove through Arg5 (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999). Such conserved interactions suggest that they are required for Hox proteins to recognize their DNA sites. The question arises as how to automatically, rather than manually, identify this type of essential interactions for protein-DNA binding.

MarkUs, a web server for analyzing the structural and functional properties of proteins (Fischer et al., 2011), provides a good platform for the study of key interactions at protein-DNA interface. MarkUs integrates a number of bioinformatic and biophysical tools, such as PSI-BLAST (Altschul et al., 1997), ClustalW (Larkin et al., 2007), and InterProScan (Zdobnov and Apweiler, 2001) for sequence analysis, and Skan (Petrey and Honig, 2003), DelPhi (Rocchia et al., 2002), and PredUs (Zhang et al., 2011) for structural analysis. Given a query structure with unknown function, MarkUs will

automatically perform sequence and structural analysis. It compares the results with structural homologs through an interactive web page named as “annotation map”, where users can choose to display functional features of the query and its structural homologs. The basic assumption in MarkUs is that proteins with structural similarity will share similar functions (Petrey et al., 2009). Thus, through the annotation map, MarkUs allows users to confirm or refute a certain function for their query structures.

In order to automatically analyze protein-DNA interactions, I developed a function annotation module in MarkUs. In this chapter, I will first describe the user-controllable parameters for this annotation module. Then, through MarkUs’s structural visualization program, AstexViewer (Hartshorn, 2002), I will demonstrate the way to analyze interactions between protein and DNA for a given query structure. Last, I will describe how to determine key protein-DNA interactions from MarkUs’s annotation map. All the demonstration of this protein-DNA annotation module will be illustrated through a case study on the structure of Scr-Exd-*fkh250* (Joshi et al., 2007).

## 5.2 Results

### 5.2.1 **A variety of input parameters allow user-specified analysis**

The function annotation module provides users with three input parameters to control the analysis of protein-DNA interactions (Figure 5-1). The first one is the cutoff distance to define a direct contact. By default, a residue-base contact is considered to form if two heavy atoms are within 6.0 Å from each other (Figure 5-1). The second parameter is the minimum number of base-pairs required to be existed in the input structure. The reason to include this parameter is because certain DNA helical

parameters, such as minor groove widths, require a number of consecutive base-pairs in order to be measured from structure. The default value for this parameter is five base-pairs (Figure 5-1). The third parameter is the cutoff to define if minor groove is narrow. Since minor groove width for canonical B-DNA is 5.8 Å, the default value used here is 5.0 Å, which is a much stricter criterion (Figure 5-1).

In addition to the three input parameters, the control panel also provides users with a functional description of this module and a related reference (Figure 5-1), where user can find features of this annotation module and more detailed explanations on the structural determinants in protein-DNA interactions.

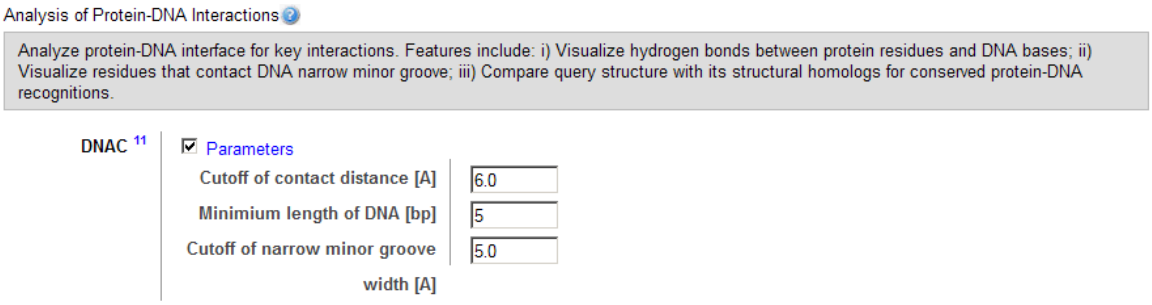


**Figure 5-1. Input parameters allow user-specified analysis of protein-DNA interactions in MarkUs.**

A screenshot of the control panel for the protein-DNA annotation module. Feature description for this annotation module is highlighted in gray. Default values for the three input parameters are shown in the boxes.

### 5.2.2 Protein-DNA interactions are visualized inside atomic structure through AstexViewer

MarkUs server has originally embedded a program named AstexViewer (Hartshorn, 2002) for protein visualization. It facilitates users to selectively inspect

features of their structures. I extended this function to annotate protein-DNA structures. Both protein-DNA interactions and DNA structures can be illustrated inside AstexViewer upon users' selections (Figure 5-2). Two types of representations are available to visualize DNA structures. One is to show DNA in lines, where every DNA's heavy atom can be seen from AstexViewer (Figure 5-2). The other is to show DNA in curvature, where DNA is represented by surface with its convex regions colored by green and concaved regions colored by gray. This type of curvature representation will provide users a clear picture on the major and minor groove shape of DNA.

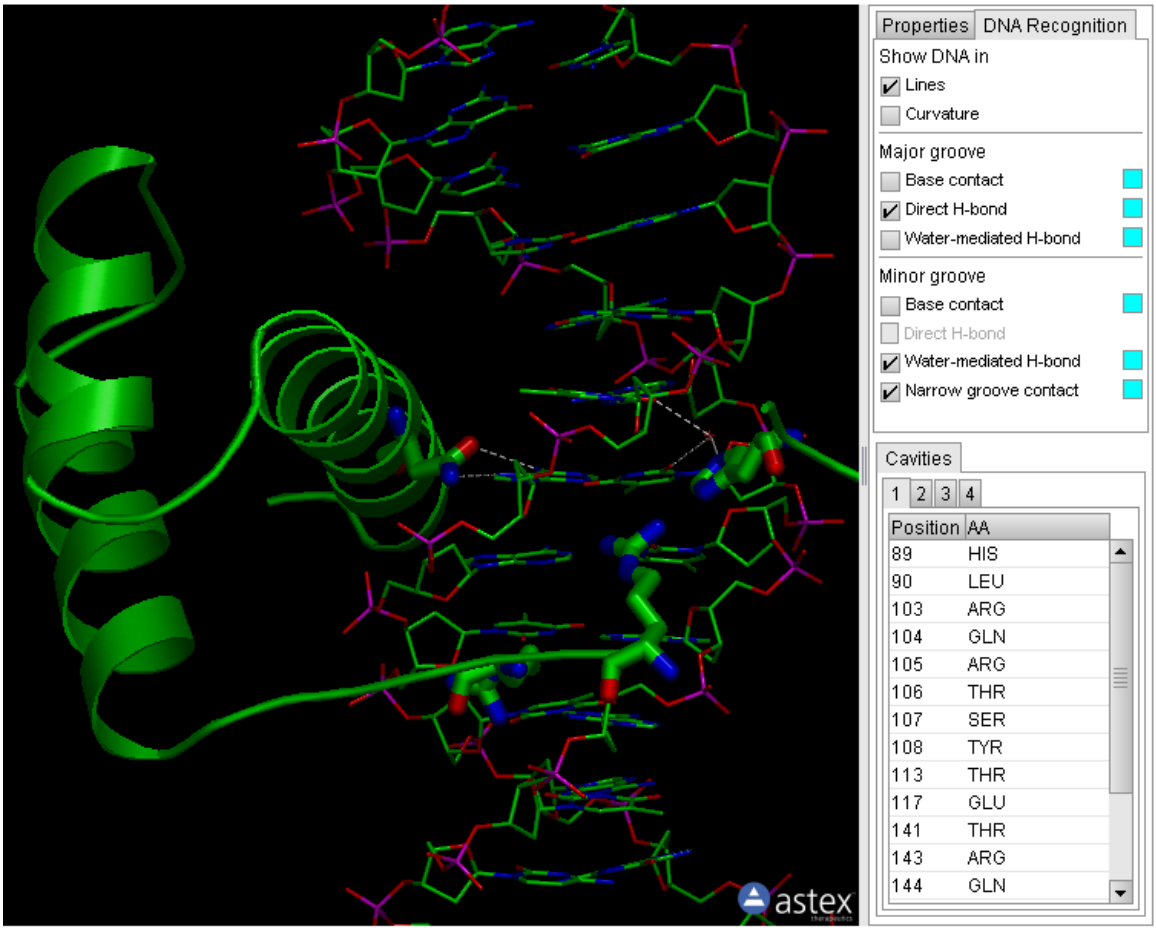


**Figure 5-2. Protein-DNA interactions are visualized inside atomic structure through AstexViewer.**

A screenshot of AstexViewer illustrating the interactions between Hox protein Scr and its specific DNA site *fkh250* (PDB ID: 2r5z (Joshi et al., 2007)). Scr's residues: His-12, Arg3, Arg5, and Asn51 are shown in sticks. The bidentate hydrogen bonds between Asn51 and an adenine in major groove together with the water-mediated hydrogen bond between His-12 and DNA's minor groove atoms are represented in dashed lines. The structure of *fkh250* is represented in lines.

The other feature of AstexViewer is that protein-DNA interactions can be visualized inside the query structure upon user's selection (Figure 5-2). These interactions are classified by DNA's major and minor groove (Figure 5-2). For each type of groove, users can select to show residues that make contacts or form direct/water-mediated hydrogen bonds with DNA base atoms (Figure 5-2) (see Method section for definitions of contact and hydrogen bond). Residues belong to each category can be highlighted with user-specified color. If no residue from the query structure was found to belong to one particular category, then the label for that category will be automatically colored by light gray and no color box will be shown (Figure 5-2). Since the recognition of narrow minor groove was found to be widely used for protein-DNA interactions (Rohs et al., 2009b), an option was added for residues belong to this category.

In the case study of Scr-Exd-*fkh250* structure, AstexViewer shows that Scr's Asn51 make bidentate hydrogen bonds with an adenine in the major groove, and His-12, Arg3, and Arg5 contact DNA's narrow minor groove (Figure 5-2). In addition, AstexViewer also shows that His-12 forms two water-mediated hydrogen bonds with DNA bases in the minor groove (Figure 5-2).

### 5.2.3 **Conserved protein-DNA interactions are identified from an annotation map**

Given many protein-DNA interactions shown in AstexViewer, the question arises as which one(s) is more important. The answer can be found by comparing the query protein with its structural homologs. The assumption is that a highly conserved interaction must be essential for protein-DNA binding.

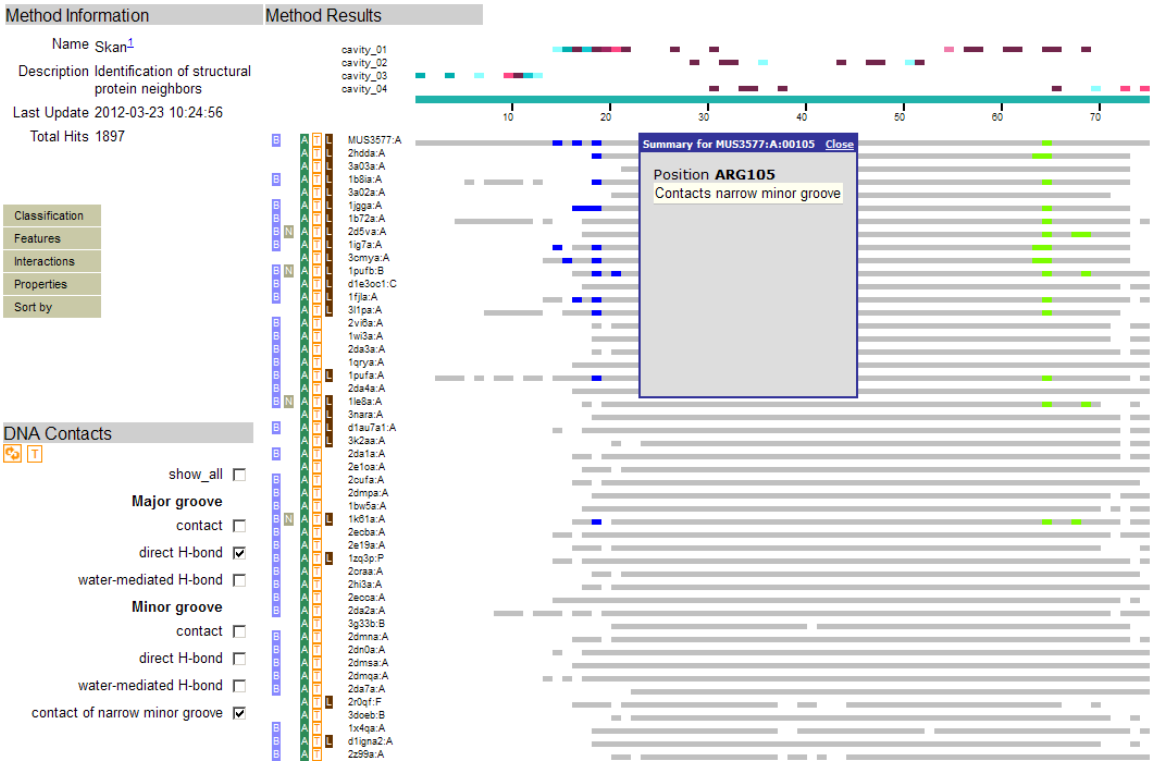


**Figure 5-3. Conserved protein-DNA interactions are identified from an annotation map.**

A screenshot of an annotation map for Hox protein Scr from the structure of Scr-Exd-*fkh250* complex (PDB ID: 2r5z (Joshi et al., 2007)). For Scr and its structural homologs, residues that contact DNA narrow minor grooves are highlighted in dark blue and residues that make direct hydrogen bonds with DNA bases are highlighted in light green. The pop-up box shows the DNA-contacting summary for Scr's Arg5.

The annotation map in MarkUs provides a convenient way to visualize the degree of conservation for certain bioinformatic or biophysical feature. I expanded this function

to protein-DNA complexes for identifying conserved interactions (Figure 5-3). On the annotation map, users can selectively highlight residues belong to each recognition category as described in the previous section (Figure 5-3). These residues will be colored on both the query and its structural homologs in an alignment map, where users can identify conserved protein-DNA interactions directly. For example, Hox protein Scr's Arg5 contacts DNA narrow minor groove and its Asn51 makes direct hydrogen bond with DNA base in the major groove. Both two interactions are conserved in many Scr's structure homologs as revealed from the annotation map (Figure 5-3).

Another feature of the annotation map is that if user moves the mouse onto one of the highlighted residues, a box will pop up to describe the DNA-interacting properties of this residue, e.g. Scr's Arg5 contacts narrow minor groove (Figure 5-3). In this way, users can quickly spot the identity and function of this residue.

## 5.3 Discussion

### 5.3.1 **An automatic way to analyze protein-DNA interactions**

I have described a function annotation module in MarkUs for analyzing protein-DNA interactions. Starting from an atomic structure of protein-DNA complexes, this module employs a robust program to analyze protein-DNA interactions. It automatically characterizes residues that recognize DNA through direct contacts, hydrogen bonds, or narrow minor groove shape readout. Upon users' selections, residues belong to each category will be displayed inside the query structure, therefore providing a straightforward way to illustrate protein-DNA interactions. In addition, AstexViewer can

also display the curvature of DNA structure, which will be helpful for users to study DNA shape readout.

The other feature of this annotation module is identifying essential protein-DNA interactions from the comparison of query with its structural homologs. This is facilitated by a database of pre-compiled structures and an annotation map in MarkUs. Every protein-DNA structure in the database has its DNA-contacting residues characterized. All these residues were stored in a database and displayed in the annotation map if they belong to query protein's structural homologs. Residues with different functions are represented by different colors. Therefore, users can identify conserved protein-DNA interactions easily from the annotation map.

### 5.3.2 **Integration with high-throughput approaches to identify genomic binding sites for transcription factors**

The function annotation module in MarkUs provides an automatic and convenient way to analyze protein-DNA interactions. Essential interactions can be identified from the comparison of query and its structure homologs. This knowledge will be useful for studying query protein's recognition to a number of different DNA sites. Through the homology modeling method described in the previous chapter, a structural model can be built by superposing DNA base-pairs that engage in conserved interactions. Evaluation of homology models will provide insights on the preference of query protein to various DNA sites.

Homology models built based on knowledge obtained from this annotation module can also help to refine *in vivo* binding sites for transcription factors (TFs). Recent

advances in high-throughput measurements, such as ChIP-chip (Ren et al., 2000), ChIP-seq (Johnson et al., 2007; Park, 2009), and ChIP-exo (Rhee and Pugh, 2011), not only have uncovered genomic binding sites for a large number of TFs, but also boosted our knowledge on transcriptional regulatory networks. However, due to cofactor-mediated DNA binding and DNA looping, false positive DNA sites perplex measurements from high-throughput approaches. Homology modeling provides an effective way to refine genomic binding sites when there is a template TF-DNA structure available. For each potential DNA site, a structural model can be built and evaluated to infer the likelihood of binding. The integration of homology modeling with high-throughput approaches will increase the accuracy of identifying TF's genomic binding sites and enhance our understanding on transcriptional regulation.

## 5.4 Method

### 5.4.1 **Analysis of protein-DNA interface**

A program written in Python is used to analyze protein-DNA interface. A residue is considered to contact DNA's major or minor groove if it has a heavy atom that is within 6.0 Å to any heavy atoms of DNA bases. Direct and water-mediated hydrogen bond is defined by HBPlus (McDonald and Thornton, 1994). DNA minor groove width is measured from atomic structure by Curves (Lavery and Sklenar, 1989). DNA base-pairings are obtained from 3DNA (Lu and Olson, 2008).

### 5.4.2 **Protein-DNA annotation module in MarkUs**

The functional annotation server MarkUs can be visited at the following link: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Mark-Us. It contains

a database of pre-compiled structures, with which the query can be compared on the annotation map. Every protein-DNA structure in this database was analyzed by the method described above and the analysis results were stored in MarkUs's MySQL database.

Three new tables were added to MarkUs's MySQL database for the protein-DNA annotation module. A table named "dnac_run" contains parameters for analyzing protein-DNA interface. Another table called "dnac_sidechain" includes information of residues that contact DNA's major, minor, and narrow minor grooves. The third table titled "dnac_hbond" stores information of direct and water-mediated hydrogen bonds between protein residues and DNA bases.

# Chapter 6. Conclusions

## 6.1 Significance of research

Protein-DNA specific recognition is essential for the normal functions of biological systems. It ensures the successful execution of transcription, replication, storage, repair, and recombination both spatially and temporally (Dynlacht, 1997; Vaquerizas et al., 2009). In terms of three-dimensional structures, the specificity of protein-DNA complexes origins from base readout, referring to interactions between protein side-chains and DNA bases, and from shape readout, defined as the recognition of locally or globally deformed DNA structure (Rohs et al., 2010). The *Drosophila* Hox proteins, which are eight homeodomain-containing transcription factors determining body plan on the anterior-posterior axis, were found to recognize their binding sites mainly through base readout in the major groove (Gehring et al., 1994a). However, homeodomains between Hox proteins are highly similar both in terms of sequences and structures, raising the question of how Hox proteins achieve their DNA-binding specificities.

In this dissertation, the above question is answered from three major aspects: i) what is the shared feature or difference between binding sites for Hox proteins; ii) how does Hox protein recognizes such feature; and iii) on a general perspective, given an atomic structure of protein-DNA complex, how to automatically identify the essential interactions.

### 6.1.1 **Latent specificity evoked by cofactors: anterior shape vs. posterior shape**

The SELEX-seq experiments provide a large resource of binding sites for all eight *Drosophila* Hox proteins in complex with their cofactors Exd and Hth (simply referered as Exd) (Slattery et al., 2011b). DNA structures predicted for high-affinity SELEX-seq sequences by MC simulations uncovered different preferences of DNA minor groove between anterior and posterior Hox proteins (Slattery et al., 2011b). The four anterior Hox proteins: Lab, Pb, Dfd, and Scr, preferentially bind to DNA with two narrow regions in the core motif, whereas the four posterior Hox proteins: Antp, Ubx, AbdA, and AbdB specifically recognize DNA with only one narrow region in the DNA minor groove. Furthermore, a knowledge-based method built on thousands of MC simulations suggested a similar observation on a much larger set of SELEX-seq sequences that have relative affinity higher than 0.1 to Hox-Exd complexes.

The finding of DNA local shape preferences extends previous study on the structural determinant of Scr's specificity (Joshi et al., 2007) to all eight *Drosophila* Hox proteins. The extra narrow region in Scr's *in vivo* binding site *fkh250* was found to account for Scr's specificity over a non-specific site $fkh250^{con}$ (Joshi et al., 2007). Now, the MC simulations suggest that, not only Scr, but all the eight Hox proteins have specific preference of DNA minor groove width (Slattery et al., 2011b). Moreover, this type of shape preference is evoked by Exd, because the minor groove contacting element is Hox protein's N-terminal, which is very flexible when Hox binds to DNA alone (Fraenkel and Pabo, 1998; Gehring et al., 1994a), but becomes stabilized when Hox binds to DNA together with Exd (Joshi et al., 2007; LaRonde-LeBlanc and Wolberger, 2003; Passner et al., 1999; Piper et al., 1999). It is the interaction between Exd and Hox's YPWM motif

help to position Hox's N-terminal arm into DNA minor groove to recognize specific DNA local shape.

### 6.1.2 **Anterior shape is recognized by Arg3**

Given the two types of DNA minor groove shape preferred by Hox proteins, the question becomes how Hox protein recognizes such shape. To answer this question, a homology modeling method was developed to infer the minor groove recognition mode of Hox protein Scr to its high affinity binding sites. Homology models suggested that Arg3 from Scr's N-terminal arm is used to recognize the extra narrow minor groove for sequences with core motifs: 5'-ATTAAT-3', 5'-ATTGAT-3', and 5'-ATAAAT-3'. In contrast, for sequences with posterior shape represented by the core motif 5'-ATTTAT-3', Arg3 is not employed to bind their minor grooves.

The minor groove recognition modes inferred from homology models agree with previous observations from the Exd-Scr-*fkh250* and Exd-Scr-*fkh250*$^{con}$ structures (Joshi et al., 2007). It also extends the Exd-Scr-*fkh250* recognition to two other core motifs: 5'-ATTGAT-3' and 5'-ATAAAT-3'. Homology models provide a useful approach to infer the recognition mode for these two types of sequences, because no Scr-Exd-DNA structure is available for these two motifs so far. From the methodology perspective, although there are several homology modeling methods have been developed and applied to study protein-DNA recognition (Alibes et al., 2010; Ashworth et al., 2006; Gao and Skolnick, 2009; Morozov and Siggia, 2007; Siggers and Honig, 2007; Thyme et al., 2009), no method has so far focused on minor groove shape readout. Our method built on a simple implicit solvent model, rapid backbone sampling, and extensive side-chain

prediction provides a robust framework to build accurate homology models to study protein-DNA minor groove recognition.

### 6.1.3 **An automatic server to analyze protein-DNA interactions**

Structural studies on Hox-cofactor-DNA complexes intrigued us to develop an automatic approach for analyzing protein-DNA interactions. Facilitated by the protein function annotation server MarkUs, we implemented an analysis module to automatically elucidate protein-DNA interactions. Through a structural visualization program named AstexViewer, various types of protein-DNA interactions can be displayed inside query structure. Furthermore, all these interactions can be compared with query's structural homologs, thereby illustrating the degree of conservation of each protein-DNA interaction to users.

Structural knowledge obtained from this annotation module will be useful for building homology models of protein-DNA complexes. Thus, through evaluating homology models, the preference of transcription factor's binding sites can be inferred. In this way, high-throughput measurements on transcription factors' genomic binding sites can be further refined by identifying potential false positive sites.

## 6.2 Future directions

The combination of MC simulation and homology modeling offers a new direction to investigate structural determinants of protein-DNA specific interactions. Along this road, several questions appear to be interesting to ask.

### 6.2.1 **Are there any other factors determining Hox specificity?**

The first question would be, in addition to minor groove shape, if there is any other structural determinant for Hox-DNA recognition. The reason to ask this question is because some sequence motifs from SELEX-seq have anterior shape, but are not preferred by all the four anterior Hox proteins (Slattery et al., 2011b). For example, the yellow motif 5'-ATGGAT-3' is only preferentially bound to Lab, but not to Pb, Dfd, or Scr. There must be some determinants, other than minor groove shape, play a role. To answer this question, homology models of anterior Hox and sequences containing the yellow motif can be built, followed by structural analysis to find out if there are any favorite interactions between the yellow motif and Lab, but not observed between yellow motif and the other three anterior Hox proteins. In addition to homology models, crystallographic and NMR structures would also be good ways to find out other determinants for Hox specificity.

The Hox cofactor, Exd, has been demonstrated to be a crucial element to evoke the latent specificity for Hox proteins (Slattery et al., 2011b). Another Hox cofactor, Hth, is required for the nuclear localization of Hox (Rieckhof et al., 1997), but its role in determining Hox DNA-binding specificity still remains elusive. SELEX-seq can be performed for full length heterotrimer Hth-Exd-Hox to identify their DNA sites, where potential structural determinants could be inferred. Since the heterotrimer binds to a much longer DNA sites than Exd-Hox does, it would be time-consuming for MC simulation method to study the intrinsic shape of DNA-binding sites. Rapid coarse-grained model (Chen et al., 2010) would be a good choice to solve this problem. Again, crystallographic and NMR structures of heterotrimer-DNA will provide detailed

information about protein-DNA contacts and protein-protein contact in order for us to understand the origins of heterotrimer's specificity.

### 6.2.2 **Specificity determinants for other transcription factor families**

The same approach on investigating Hox specificity can be applied to other transcription factor families. Several high-throughput experiments have been performed to determine the binding sites for many transcription factor families (Grove et al., 2009; Noyes et al., 2008b; Siggers et al., 2011). The question unanswered is what the structural determinants of specificity are for each family. A systematic catalog needs to be constructed to annotate each transcription factor's preferential DNA-binding sites, structural information, and implications to human diseases. Such catalog will help to compare the DNA sites preferences between members within the same family. Potential structural determinants for specificity could be uncovered from available structures, which are either experimentally determined or obtained from homology modeling. Furthermore, this comprehensive database will provide possible explanations on the molecular basis for known mutations and therefore help clinical researchers to design therapeutic strategy to treat human diseases.

The protein-DNA annotation module in MarkUs can be further developed to bring the knowledge of protein-DNA recognition to a much wider audience. The server can include the function that makes predictions on the free DNA structures and builds homology models of protein-DNA complexes. In this way, it will not allow researchers to analyze available structural information, but also let them infer potential specificity determinants from computational models. In addition, connecting protein-DNA

complexes with genome-wide study and clinical research is also a promising direction. More functional module could be introduced to MarkUs to integrate research results from genome-wide studies, such as high-throughput measurements, network mapping. Combining knowledge on the structural level with genomic research will elucidate the biological systems at a more comprehensive perspective. Also, integrating structural information with clinical studies will provide researchers a much clearer picture on the mechanisms for human diseases.

# Bibliography

Abate-Shen, C. (2002). Deregulated homeobox gene expression in cancer: cause or consequence? Nat Rev Cancer *2*, 777-785.

Affolter, M., Slattery, M., and Mann, R.S. (2008). A lexicon for homeodomain-DNA recognition. Cell *133*, 1133-1135.

Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science *242*, 899-907.

Alexander, T., Nolte, C., and Krumlauf, R. (2009). Hox genes and segmentation of the hindbrain and axial skeleton. Annu Rev Cell Dev Biol *25*, 431-456.

Alibes, A., Nadra, A.D., De Masi, F., Bulyk, M.L., Serrano, L., and Stricher, F. (2010). Using protein design algorithms to understand the molecular basis of disease caused by protein-DNA interactions: the Pax6 example. Nucleic Acids Res.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. Nucleic acids research *36*, D419-425.

Andrews, A.J., and Luger, K. (2011). Nucleosome Structure(s) and Stability: Variations on a Theme. In Annual Review of Biophysics, Vol 40, D.C.D.K.A.W.J.R. Rees, ed., pp. 99-117.

Arya, G., and Schlick, T. (2006). Role of histone tails in chromatin folding revealed by a mesoscopic oligonucleosome model. Proceedings of the National Academy of Sciences of the United States of America *103*, 16236-16241.

Ashworth, J., Havranek, J.J., Duarte, C.M., Sussman, D., Monnat, R.J., Jr., Stoddard, B.L., and Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. Nature *441*, 656-659.

Baburajendran, N., Jauch, R., Tan, C.Y.Z., Narasimhan, K., and Kolatkar, P.R. (2011). Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers. Nucleic Acids Research *39*, 8213-8222.

Bartfeld, D., Shimon, L., Couture, G.C., Rabinovich, D., Frolow, F., Levanon, D., Groner, Y., and Shakked, Z. (2002). DNA recognition by the RUNX1 transcription factor is

mediated by an allosteric transition in the RUNT domain and by DNA bending. Structure *10*, 1395-1407.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., *et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc *4*, 393-411.

Berthelsen, J., Zappavigna, V., Mavilio, F., and Blasi, F. (1998). Prep1, a novel functional partner of Pbx proteins. Embo J *17*, 1423-1433.

Bishop, E.P., Rohs, R., Parker, S.C.J., West, S.M., Liu, P., Mann, R.S., Honig, B., and Tullius, T.D. (2011). A Map of Minor Groove Shape and Electrostatic Potential from Hydroxyl Radical Cleavage Patterns of DNA. Acs Chemical Biology *6*, 1314-1320.

Blanco, F.J., and Montoya, G. (2011). Transient DNA/RNA-protein interactions. Febs Journal *278*, 1643-1650.

Bradley, P.L., Haberman, A.S., and Andrew, D.J. (2001). Organ formation in Drosophila: specification and morphogenesis of the salivary gland. Bioessays *23*, 901-911.

Bulyk, M.L. (2003). Computational prediction of transcription-factor binding site locations. Genome Biol *5*, 201.

Bulyk, M.L., Huang, X., Choo, Y., and Church, G.M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proceedings of the National Academy of Sciences of the United States of America *98*, 7158-7163.

Cahill, S., Cahill, M., and Cahill, K. (2003). On the kinematics of protein folding. Journal of Computational Chemistry *24*, 1364-1370.

Campbell, C.T., and Kim, G. (2007). SPR microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics. Biomaterials *28*, 2380-2392.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., *et al.* (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell *18*, 662-674.

Capovilla, M., and Botas, J. (1998). Functional dominance among Hox genes: repression dominates activation in the regulation of Dpp. Development *125*, 4949-4957.

Casares, F., Calleja, M., and Sanchez-Herrero, E. (1996). Functional similarity in appendage specification by the Ultrabithorax and abdominal-A Drosophila HOX genes. Embo J *15*, 3934-3942.

Chang, C.P., Brocchieri, L., Shen, W.F., Largman, C., and Cleary, M.L. (1996). Pbx modulation of Hox homeodomain amino-terminal arms establishes different DNA-binding specificities across the Hox locus. Mol Cell Biol *16*, 1734-1745.

Cheatham, T.E., 3rd (2004). Simulation and modeling of nucleic acid structure, dynamics and interactions. Curr Opin Struc Biol *14*, 360-367.

Cheatham, T.E., 3rd, and Kollman, P.A. (2000). Molecular dynamics simulation of nucleic acids. Annu Rev Phys Chem *51*, 435-471.

Chen, J., Darst, S.A., and Thirumalai, D. (2010). Promoter melting triggered by bacterial RNA polymerase occurs in three steps. Proceedings of the National Academy of Sciences of the United States of America *107*, 12523-12528.

Christensen, R.G., Gupta, A., Zuo, Z., Schriefer, L.A., Wolfe, S.A., and Stormo, G.D. (2011). A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic Acids Research *39*.

Coiffier, D., Charroux, B., and Kerridge, S. (2008). Common functions of central and posterior Hox genes for the repression of head in the trunk of Drosophila. Development *135*, 291-300.

Cordeiro, T.N., Schmidt, H., Madrid, C., Juarez, A., Bernado, P., Griesinger, C., Garcia, J., and Pons, M. (2011). Indirect DNA Readout by an H-NS Related Protein: Structure of the DNA Complex of the C-Terminal Domain of Ler. Plos Pathogens *7*.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.J., Ferguson, D.M., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P.A. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. J Am Chem Soc *117*, 5179-5197.

Crickmore, M.A., and Mann, R.S. (2008). The control of size in animals: insights from selector genes. Bioessays *30*, 843-853.

Czapla, L., Swigon, D., and Olson, W.K. (2008). Effects of the nucleoid protein HU on the structure, flexibility, and ring-closure properties of DNA deduced from Monte Carlo simulations. Journal of molecular biology *382*, 353-370.

De Masi, F., Grove, C.A., Vedenko, A., Alibes, A., Gisselbrecht, S.S., Serrano, L., Bulyk, M.L., and Walhout, A.J.M. (2011). Using a structural and logics systems approach to infer bHLH-DNA binding specificity determinants. Nucleic Acids Research *39*, 4553-4563.

Dobrovolskaia, I.V., Kenward, M., and Arya, G. (2010). Twist Propagation in Dinucleosome Arrays. Biophys J *99*, 3355-3364.

Dynlacht, B.D. (1997). Regulation of transcription by proteins that control the cell cycle. Nature *389*, 149-152.

Ekker, S.C., Jackson, D.G., von Kessler, D.P., Sun, B.I., Young, K.E., and Beachy, P.A. (1994). The degree of variation in DNA sequence recognition among four Drosophila homeotic proteins. Embo J *13*, 3551-3560.

Ekker, S.C., von Kessler, D.P., and Beachy, P.A. (1992). Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. Embo J *11*, 4059-4072.

Ellenberger, T. (1994). Getting a Grip on DNA Recognition - Structures of the Basic Region Leucine-Zipper, and the Basic Region Helix-Loop-Helix DNA-Binding Domains. Curr Opin Struc Biol *4*, 12-21.

Elrod-Erickson, M., Rould, M.A., Nekludova, L., and Pabo, C.O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure *4*, 1171-1180.

Fischer, M., Zhang, Q.C., Dey, F., Chen, B.Y., Honig, B., and Petrey, D. (2011). MarkUs: a server to navigate sequence-structure-function space. Nucleic Acids Res *39*, W357-361.

Fraenkel, E., and Pabo, C.O. (1998). Comparison of X-ray and NMR structures for the Antennapedia homeodomain-DNA complex. Nat Struct Biol *5*, 692-697.

Fujii, S., Kono, H., Takenaka, S., Go, N., and Sarai, A. (2007). Sequence-dependent DNA deformability studied using molecular dynamics simulations. Nucleic acids research *35*, 6063-6074.

Fuxreiter, M., Simon, I., and Bondos, S. (2011). Dynamic protein-DNA recognition: beyond what can be seen. Trends in Biochemical Sciences *36*, 415-423.

Gao, M., and Skolnick, J. (2009). A threading-based method for the prediction of DNA-binding proteins with application to the human genome. PLoS Comput Biol *5*, e1000567.

Garvie, C.W., and Wolberger, C. (2001). Recognition of specific DNA sequences. Mol Cell *8*, 937-946.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., *et al.* (2010). A map of open chromatin in human pancreatic islets. Nat Genet *42*, 255-259.

Gebelein, B., Culi, J., Ryoo, H.D., Zhang, W., and Mann, R.S. (2002). Specificity of Distalless repression and limb primordia development by abdominal Hox proteins. Dev Cell *3*, 487-498.

Gebelein, B., McKay, D.J., and Mann, R.S. (2004). Direct integration of Hox and segmentation gene inputs during Drosophila development. Nature *431*, 653-659.

Gehring, W.J., Affolter, M., and Burglin, T. (1994a). Homeodomain proteins. Annual review of biochemistry *63*, 487-526.

Gehring, W.J., Qian, Y.Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A.F., Resendez-Perez, D., Affolter, M., Otting, G., and Wuthrich, K. (1994b). Homeodomain-DNA recognition. Cell *78*, 211-223.

Ghannam, G., Takeda, A., Camarata, T., Moore, M.A., Viale, A., and Yaseen, N.R. (2004). The oncogene Nup98-HOXA9 induces gene transcription in myeloid cells. J Biol Chem *279*, 866-875.

Ghosh, G., van Duyne, G., Ghosh, S., and Sigler, P.B. (1995). Structure of NF-kappa B p50 homodimer bound to a kappa B site. Nature *373*, 303-310.

Goodman, F.R. (2002). Limb malformations and the human HOX genes. Am J Med Genet *112*, 256-265.

Goodman, F.R., Bacchelli, C., Brady, A.F., Brueton, L.A., Fryns, J.P., Mortlock, D.P., Innis, J.W., Holmes, L.B., Donnenfeld, A.E., Feingold, M., *et al.* (2000). Novel HOXA13 mutations and the phenotypic spectrum of hand-foot-genital syndrome. Am J Hum Genet *67*, 197-202.

Goodman, F.R., and Scambler, P.J. (2001). Human HOX gene mutations. Clin Genet *59*, 1-11.

Gordon, B.R.G., Li, Y., Cote, A., Weirauch, M.T., Ding, P., Hughes, T.R., Navarre, W.W., Xia, B., and Liu, J. (2011). Structural basis for recognition of AT-rich DNA by unrelated xenogeneic silencing proteins. Proceedings of the National Academy of Sciences of the United States of America *108*, 10690-10695.

Grove, C.A., De Masi, F., Barrasa, M.I., Newburger, D.E., Alkema, M.J., Bulyk, M.L., and Walhout, A.J. (2009). A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell *138*, 314-327.

Guertin, M.J., and Lis, J.T. (2010). Chromatin landscape dictates HSF binding to target DNA elements. PLoS Genet *6*.

Guzikevich-Guerstein, G., and Shakked, Z. (1996). A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. Nat Struct Biol *3*, 32-37.

Hartshorn, M.J. (2002). AstexViewer: a visualisation aid for structure-based drug design. Journal of computer-aided molecular design *16*, 871-881.

Havranek, J.J., Duarte, C.M., and Baker, D. (2004). A simple physical model for the prediction and design of protein-DNA interactions. J Mol Biol *344*, 59-70.

Hegde, R.S. (2002). The papillomavirus E2 proteins: structure, function, and biology. Annu Rev Biophys Biomol Struct *31*, 343-360.

Hersh, B.M., and Carroll, S.B. (2005). Direct regulation of knot gene expression by Ultrabithorax and the evolution of cis-regulatory elements in Drosophila. Development *132*, 1567-1577.

Honig, B., and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. Science *268*, 1144-1149.

Honig, B., and Rohs, R. (2011). BIOPHYSICS Flipping Watson and Crick. Nature *470*, 472-473.

Hueber, S.D., Weiller, G.F., Djordjevic, M.A., and Frickey, T. (2010). Improving Hox protein classification across the major model organisms. PLoS One *5*, e10820.

Huth, J.R., Bewley, C.A., Nissen, M.S., Evans, J.N., Reeves, R., Gronenborn, A.M., and Clore, G.M. (1997). The solution structure of an HMG-I(Y)-DNA complex defines a new architectural minor groove binding motif. Nat Struct Biol *4*, 657-665.

Jacobson, M.P., Friesner, R.A., Xiang, Z., and Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol *320*, 597-608.

Jayaram, B., Sharp, K.A., and Honig, B. (1989). The electrostatic potential of B-DNA. Biopolymers *28*, 975-993.

Johnson, A.D. (1995). Molecular mechanisms of cell-type determination in budding yeast. Curr Opin Genet Dev *5*, 552-558.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530-543.

Joshi, R., Sun, L., and Mann, R. (2010). Dissecting the functional specificities of two Hox proteins. Genes Dev *24*, 1533-1545.

Jurgens, G., Wieschaus, E., Nussleinvolhard, C., and Kluding, H. (1984). Mutations Affecting the Pattern of the Larval Cuticle in Drosophila-Melanogaster .2. Zygotic Loci on the 3rd Chromosome. Roux Arch Dev Biol *193*, 283-295.

Kalodimos, C.G., Boelens, R., and Kaptein, R. (2004). Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. Chem Rev *104*, 3567-3586.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., *et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Kaplan, T., Li, X.Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet *7*, e1001290.

Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993a). Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature *365*, 520-527.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993b). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512-520.

Kitayner, M., Rozenberg, H., Rohs, R., Suad, O., Rabinovich, D., Honig, B., and Shakked, Z. (2010). Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. Nat Struct Mol Biol *17*, 423-429.

Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778-795.

Lane, W.J., and Darst, S.A. (2006). The structural basis for promoter -35 element recognition by the group IV sigma factors. PLoS Biol *4*, e269.

Lankas, F., Spackova, N., Moakher, M., Enkhbayar, P., and Sponer, J. (2010). A measure of bending in nucleic acids structures applied to A-tract DNA. Nucleic Acids Research *38*, 3414-3422.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

LaRonde-LeBlanc, N.A., and Wolberger, C. (2003). Structure of HoxA9 and Pbx1 bound to DNA: Hox hexapeptide and DNA recognition anterior to posterior. Genes Dev *17*, 2060-2072.

Lavery, R., and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. J Biomol Struct Dyn *6*, 655-667.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Lewis, E.B. (1992). The 1991 Albert Lasker Medical Awards. Clusters of master control genes regulate the development of higher organisms. Jama *267*, 1524-1531.

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G., and Lu, P. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science *271*, 1247-1254.

Li, J.Y., Cui, Z., Gao, X.F., Jiang, J.Y., Jing, T.L., Shi, L., Peng, Y.H., and Sun, Y.H. (2011). Re: Markus J. Bader, Christian Gratzke, Michael Seitz, et al. The "all-seeing needle": initial results of an optical puncture system confirming access in percutaneous nephrolithotomy. Eur Urol 2011;59:1054-9. European urology *60*, e42-43; author reply e44.

Li, X., and McGinnis, W. (1999). Activity regulation of Hox proteins, a mechanism for altering functional specificity in development and evolution. Proc Natl Acad Sci U S A *96*, 6802-6807.

Lindemose, S., Nielsen, P.E., Hansen, M., and Mollegaard, N.E. (2011). A DNA minor groove electronegative potential genome map based on photo-chemical probing. Nucleic Acids Research *39*, 6269-6276.

Lohmann, I., McGinnis, N., Bodmer, M., and McGinnis, W. (2002). The Drosophila Hox gene deformed sculpts head morphology via direct regulation of the apoptosis activator reaper. Cell *110*, 457-466.

Lu, X.J., and Olson, W.K. (2008). 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc *3*, 1213-1227.

Lufkin, T., Dierich, A., LeMeur, M., Mark, M., and Chambon, P. (1991). Disruption of the Hox-1.6 homeobox gene results in defects in a region corresponding to its rostral domain of expression. Cell *66*, 1105-1119.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic acids research *29*, 2860-2874.

Maerkl, S.J., and Quake, S.R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. Science *315*, 233-237.

Mann, R.S. (1995). The specificity of homeotic gene function. Bioessays *17*, 855-863.

Mann, R.S., and Carroll, S.B. (2002). Molecular mechanisms of selector gene function and evolution. Curr Opin Genet Dev *12*, 592-600.

Mann, R.S., Lelli, K.M., and Joshi, R. (2009). Hox specificity unique roles for cofactors and collaborators. Curr Top Dev Biol *88*, 63-101.

McDonald, I.K., and Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. J Mol Biol *238*, 777-793.

McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell *68*, 283-302.

McKay, D.B., and Steitz, T.A. (1981). Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. Nature *290*, 744-749.

Meinke, G., and Sigler, P.B. (1999). DNA-binding mechanism of the monomeric orphan nuclear receptor NGFI-B. Nat Struct Biol *6*, 471-477.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Merabet, S., Saadaoui, M., Sambrani, N., Hudry, B., Pradel, J., Affolter, M., and Graba, Y. (2007). A unique Extradenticle recruitment mode in the Drosophila Hox protein Ultrabithorax. Proceedings of the National Academy of Sciences of the United States of America *104*, 16946-16951.

Milech, N., Kees, U.R., and Watt, P.M. (2001). Novel alternative PBX3 isoforms in leukemia cells with distinct interaction specificities. Genes Chromosomes Cancer *32*, 275-280.

Miotto, B., and Graba, Y. (2010). Control of DNA replication: A new facet of Hox proteins? Bioessays *32*, 800-807.

Mollegaard, N.E., Lindemose, S., and Nielsen, P.E. (2005). Uranyl photoprobing of nonbent A/T- and bent A-tracts. A difference of flexibility? Biochemistry *44*, 7855-7863.

Morozov, A.V., Havranek, J.J., Baker, D., and Siggia, E.D. (2005). Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res *33*, 5781-5798.

Morozov, A.V., and Siggia, E.D. (2007). Connecting protein structure with predictions of regulatory sites. Proc Natl Acad Sci U S A *104*, 7068-7073.

Mrinal, N., Tomar, A., and Nagaraju, J. (2011). Role of sequence encoded kappa B DNA geometry in gene regulation by Dorsal. Nucleic Acids Research *39*, 9574-9591.

Muragaki, Y., Mundlos, S., Upton, J., and Olsen, B.R. (1996). Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science *272*, 548-551.

Naiche, L.A., Harrelson, Z., Kelly, R.G., and Papaioannou, V.E. (2005). T-box genes in vertebrate development. Annu Rev Genet *39*, 219-239.

Nair, S.K., and Burley, S.K. (2003). X-ray structures of Myc-Max and Mad-Max recognizing DNA. Molecular bases of regulation by proto-oncogenic transcription factors. Cell *112*, 193-205.

Natoli, G. (2010). Maintaining cell identity through global control of genomic organization. Immunity *33*, 12-24.

Nikolova, E.N., Kim, E., Wise, A.A., O'Brien, P.J., Andricioaei, I., and Al-Hashimi, H.M. (2011). Transient Hoogsteen base pairs in canonical duplex DNA. Nature *470*, 498-U484.

Noro, B., Culi, J., McKay, D.J., Zhang, W., and Mann, R.S. (2006). Distinct functions of homeodomain-containing and homeodomain-less isoforms encoded by homothorax. Genes Dev *20*, 1636-1650.

Noro, B., Lelli, K., Sun, L., and Mann, R.S. (2011). Competition for cofactor-dependent DNA binding underlies Hox phenotypic suppression. Genes Dev *25*, 2327-2332.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008a). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

Noyes, M.B., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M.H., and Wolfe, S.A. (2008b). A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. Nucleic acids research *36*, 2547-2560.

Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y., and Matthews, B.W. (1982). The molecular basis of DNA-protein recognition inferred from the structure of cro repressor. Nature *298*, 718-723.

Olson, W.K., and Zhurkin, V.B. (2011). Working the kinks out of nucleosomal DNA. Curr Opin Struc Biol *21*, 348-357.

Orozco, M., Noy, A., and Perez, A. (2008). Recent advances in the study of nucleic acid flexibility by molecular dynamics. Curr Opin Struc Biol *18*, 185-193.

Otting, G., Qian, Y.Q., Billeter, M., Muller, M., Affolter, M., Gehring, W.J., and Wuthrich, K. (1990). Protein--DNA contacts in the structure of a homeodomain--DNA complex determined by nuclear magnetic resonance spectroscopy in solution. Embo J *9*, 3085-3092.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321-329.

Pabo, C.O., and Lewis, M. (1982). The operator-binding domain of lambda repressor: structure and DNA recognition. Nature *298*, 443-447.

Pan, Y., and Nussinov, R. (2007). Structural basis for p53 binding-induced DNA bending. J Biol Chem *282*, 691-699.

Pan, Y., and Nussinov, R. (2008). p53-Induced DNA bending: the interplay between p53-DNA and p53-p53 interactions. J Phys Chem B *112*, 6716-6724.

Panne, D. (2008). The enhanceosome. Curr Opin Struct Biol *18*, 236-242.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.

Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D., and Margulies, E.H. (2009). Local DNA topography correlates with functional noncoding regions of the human genome. Science *324*, 389-392.

Parker, S.C.J., and Tullius, T.D. (2011). DNA shape, genetic codes, and evolution. Curr Opin Struc Biol *21*, 342-347.

Passner, J.M., Ryoo, H.D., Shen, L., Mann, R.S., and Aggarwal, A.K. (1999). Structure of a DNA-bound Ultrabithorax-Extradenticle homeodomain complex. Nature *397*, 714-719.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

Pavletich, N.P., and Pabo, C.O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science *261*, 1701-1707.

Pavlopoulos, A., and Akam, M. (2011). Hox gene Ultrabithorax regulates distinct sets of target genes at successive stages of Drosophila haltere morphogenesis. Proceedings of the National Academy of Sciences of the United States of America *108*, 2855-2860.

Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nat Rev Genet *6*, 893-904.

Peifer, M., and Wieschaus, E. (1990). Mutations in the Drosophila gene extradenticle affect the way specific homeo domain proteins regulate segmental identity. Genes Dev *4*, 1209-1223.

Perez, A., Marchan, I., Svozil, D., Sponer, J., Cheatham, T.E., 3rd, Laughton, C.A., and Orozco, M. (2007). Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J *92*, 3817-3829.

Petrey, D., Fischer, M., and Honig, B. (2009). Structural relationships among proteins with different global topologies and their implications for function annotation strategies. Proc Natl Acad Sci U S A *106*, 17377-17382.

Petrey, D., and Honig, B. (2003). GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. Methods Enzymol *374*, 492-509.

Petrey, D., and Honig, B. (2005). Protein structure prediction: inroads to biology. Mol Cell *20*, 811-819.

Piper, D.E., Batchelor, A.H., Chang, C.P., Cleary, M.L., and Wolberger, C. (1999). Structure of a HoxB1-Pbx1 heterodimer bound to DNA: role of the hexapeptide and a fourth homeodomain helix in complex formation. Cell *96*, 587-597.

Ponder, J.W., and Richards, F.M. (1987). Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology *193*, 775-791.

Qian, S., Capovilla, M., and Pirrotta, V. (1991). The bx region enhancer, a distant cis-control element of the Drosophila Ubx gene and its regulation by hunchback and other segmentation genes. Embo J *10*, 1415-1425.

Qian, Y.Q., Resendez-Perez, D., Gehring, W.J., and Wuthrich, K. (1994). The des(1-6)antennapedia homeodomain: comparison of the NMR solution structure and the DNA-binding affinity with the intact Antennapedia homeodomain. Proceedings of the National Academy of Sciences of the United States of America *91*, 4091-4095.

Rastinejad, F., Wagner, T., Zhao, Q., and Khorasanizadeh, S. (2000). Structure of the RXR-RAR DNA-binding complex on the retinoic acid response element DR1. Embo J *19*, 1045-1054.

Regulski, M., Dessain, S., McGinnis, N., and McGinnis, W. (1991). High-affinity binding sites for the Deformed protein are required for the function of an autoregulatory enhancer of the Deformed gene. Genes Dev *5*, 278-286.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell *147*, 1408-1419.

Rice, P.A., Yang, S., Mizuuchi, K., and Nash, H.A. (1996). Crystal structure of an IHF-DNA complex: a protein-induced DNA U-turn. Cell *87*, 1295-1306.

Rieckhof, G.E., Casares, F., Ryoo, H.D., Abu-Shaar, M., and Mann, R.S. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. Cell *91*, 171-183.

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., and Honig, B. (2002). Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem *23*, 128-137.

Rohs, R., Bloch, I., Sklenar, H., and Shakked, Z. (2005a). Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. Nucleic acids research *33*, 7048-7057.

Rohs, R., Etchebest, C., and Lavery, R. (1999). Unraveling proteins: a molecular mechanics study. Biophys J *76*, 2760-2768.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu Rev Biochem *79*, 233-269.

Rohs, R., Sklenar, H., and Shakked, Z. (2005b). Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. Structure *13*, 1499-1509.

Rohs, R., West, S.M., Liu, P., and Honig, B. (2009a). Nuance in the double-helix and its role in protein-DNA recognition. Curr Opin Struct Biol *19*, 171-177.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009b). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248-1253.

Ryoo, H.D., and Mann, R.S. (1999). The control of trunk Hox specificity and activity by Extradenticle. Genes Dev *13*, 1704-1716.

Ryoo, H.D., Marty, T., Casares, F., Affolter, M., and Mann, R.S. (1999). Regulation of Hox target genes by a DNA bound Homothorax/Hox/Extradenticle complex. Development *126*, 5137-5148.

Schumacher, M.A., Choi, K.Y., Zalkin, H., and Brennan, R.G. (1994). Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science *266*, 763-770.

Schwartz, T., Behlke, J., Lowenhaupt, K., Heinemann, U., and Rich, A. (2001). Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. Nat Struct Biol *8*, 761-765.

Seeliger, D., Buelens, F.P., Goette, M., de Groot, B.L., and Grubmueller, H. (2011). Towards computional specificity screening of DNA-binding proteins. Nucleic Acids Research *39*, 8281-8290.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proceedings of the National Academy of Sciences of the United States of America *73*, 804-808.

Segal, E., and Widom, J. (2009). From DNA sequence to transcriptional behaviour: a quantitative approach. Nat Rev Genet *10*, 443-456.

Shah, N., and Sukumar, S. (2010). The Hox genes and their roles in oncogenesis. Nat Rev Cancer *10*, 361-371.

Shearn, A. (1989). The ash-1, ash-2 and trithorax genes of Drosophila melanogaster are functionally related. Genetics *121*, 517-525.

Shen, A., Higgins, D.E., and Panne, D. (2009). Recognition of AT-rich DNA binding sites by the MogR repressor. Structure *17*, 769-777.

Shen, W.F., Rozenfeld, S., Lawrence, H.J., and Largman, C. (1997). The Abd-B-like Hox homeodomain proteins can be subdivided by the ability to form complexes with Pbx1a on a novel DNA target. J Biol Chem *272*, 8198-8206.

Siggers, T., Chang, A.B., Teixeira, A., Wong, D., Williams, K.J., Ahmed, B., Ragoussis, J., Udalova, I.A., Smale, S.T., and Bulyk, M.L. (2011). Principles of dimer-specific gene regulation revealed by a comprehensive characterization of NF-kappaB family DNA binding. Nat Immunol *13*, 95-102.

Siggers, T.W., and Honig, B. (2007). Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. Nucleic Acids Res *35*, 1085-1097.

Sklenar, H., Wustner, D., and Rohs, R. (2006). Using internal and collective variables in Monte Carlo simulations of nucleic acid structures: chain breakage/closure algorithm and associated Jacobians. Journal of Computational Chemistry *27*, 309-315.

Slattery, M., Ma, L., Negre, N., White, K.P., and Mann, R.S. (2011a). Genome-Wide Tissue-Specific Occupancy of the Hox Protein Ultrabithorax and Hox Cofactor Homothorax in Drosophila. PLoS One *6*.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., *et al.* (2011b). Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell *147*, 1270-1282.

Somers, W.S., and Phillips, S.E. (1992). Crystal structure of the met repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. Nature *359*, 387-393.

St Johnston, D., and Nusslein-Volhard, C. (1992). The origin of pattern and polarity in the Drosophila embryo. Cell *68*, 201-219.

Stadler, H.S., Higgins, K.M., and Capecchi, M.R. (2001). Loss of Eph-receptor expression correlates with loss of cell adhesion and chondrogenic capacity in Hoxa13 mutant limbs. Development *128*, 4177-4188.

Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. Genes Dev *24*, 814-826.

Swigon, D., Coleman, B.D., and Olson, W.K. (2006). Modeling the Lac repressor-operator assembly: the influence of DNA looping on Lac repressor conformation. Proceedings of the National Academy of Sciences of the United States of America *103*, 9879-9884.

Taghli-Lamallem, O., Gallet, A., Leroy, F., Malapert, P., Vola, C., Kerridge, S., and Fasano, L. (2007). Direct interaction between Teashirt and Sex combs reduced proteins, via Tsh's acidic domain, is essential for specifying the identity of the prothorax in Drosophila. Dev Biol *307*, 142-151.

Tahirov, T.H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., *et al.* (2001). Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Cell *104*, 755-767.

Tainer, J.A., and Cunningham, R.P. (1993). Molecular recognition in DNA-binding proteins and enzymes. Curr Opin Biotechnol *4*, 474-483.

Takahashi, O., Hamada, J., Abe, M., Hata, S., Asano, T., Takahashi, Y., Tada, M., Miyamoto, M., Kondo, S., and Moriuchi, T. (2007). Dysregulated expression of HOX and ParaHOX genes in human esophageal squamous cell carcinoma. Oncol Rep *17*, 753-760.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol.

Temiz, N.A., and Camacho, C.J. (2009). Experimentally based contact energies decode interactions responsible for protein-DNA affinity and the role of molecular waters at the binding interface. Nucleic Acids Research *37*, 4076-4088.

Thorsteinsdottir, U., Sauvageau, G., Hough, M.R., Dragowska, W., Lansdorp, P.M., Lawrence, H.J., Largman, C., and Humphries, R.K. (1997). Overexpression of HOXA10 in murine hematopoietic cells perturbs both myeloid and lymphoid differentiation and leads to acute myeloid leukemia. Mol Cell Biol *17*, 495-505.

Thyme, S.B., Jarjour, J., Takeuchi, R., Havranek, J.J., Ashworth, J., Scharenberg, A.M., Stoddard, B.L., and Baker, D. (2009). Exploitation of binding energy for catalysis and design. Nature *461*, 1300-1304.

Tietjen, J.R., Donato, L.J., Bhimisaria, D., and Ansari, A.Z. (2011). SEQUENCE-SPECIFICITY AND ENERGY LANDSCAPES OF DNA-BINDING MOLECULES. In Methods in Enzymology, Vol 497: Synthetic Biology, Methods for Part/Device Characterization and Chassis Engineering, Pt A, C. Voigt, ed., pp. 3-30.

Tihanyi, B., Vellai, T., Regos, A., Ari, E., Mueller, F., and Takacs-Vellai, K. (2010). The C. elegans Hox gene ceh-13 regulates cell migration and fusion in a non-colinear way. Implications for the early evolution of Hox clusters. Bmc Developmental Biology *10*.

Tolstorukov, M.Y., Colasanti, A.V., McCandlish, D.M., Olson, W.K., and Zhurkin, V.B. (2007). A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. Journal of molecular biology *371*, 725-738.

Vachon, G., Cohen, B., Pfeifle, C., McGuffin, M.E., Botas, J., and Cohen, S.M. (1992). Homeotic genes of the Bithorax complex repress limb development in the abdomen of the Drosophila embryo through the target gene Distal-less. Cell *71*, 437-450.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat Rev Genet *10*, 252-263.

Wang, C., Schueler-Furman, O., and Baker, D. (2005). Improved side-chain modeling for protein-protein docking. Protein Sci *14*, 1328-1339.

Wang, X., Bryant, G.O., Floer, M., Spagna, D., and Ptashne, M. (2011). An effect of DNA sequence on nucleosome occupancy and removal. Nature Structural & Molecular Biology *18*, 507-509.

Warren, C.L., Kratochvil, N.C., Hauschild, K.E., Foister, S., Brezinski, M.L., Dervan, P.B., Phillips, G.N., Jr., and Ansari, A.Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. Proceedings of the National Academy of Sciences of the United States of America *103*, 867-872.

Watkins, D., Hsiao, C., Woods, K.K., Koudelka, G.B., and Williams, L.D. (2008). P22 c2 repressor-operator complex: mechanisms of direct and indirect readout. Biochemistry *47*, 2325-2338.

Wedeen, C., Harding, K., and Levine, M. (1986). Spatial regulation of Antennapedia and bithorax gene expression by the Polycomb locus in Drosophila. Cell *44*, 739-748.

West, S.M., Rohs, R., Mann, R.S., and Honig, B. (2010). Electrostatic interactions between arginines and the minor groove in the nucleosome. J Biomol Struct Dyn *27*, 861-866.

Wolberger, C. (1996). Homeodomain interactions. Curr Opin Struct Biol *6*, 62-68.

Wolberger, C., Dong, Y.C., Ptashne, M., and Harrison, S.C. (1988). Structure of a phage 434 Cro/DNA complex. Nature *335*, 789-795.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct *29*, 183-212.

Xiang, Z., and Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol *311*, 421-430.

Xiang, Z., Steinbach, P.J., Jacobson, M.P., Friesner, R.A., and Honig, B. (2007). Prediction of side-chain conformations on protein surfaces. Proteins *66*, 814-823.

Yang, W., and Steitz, T.A. (1995). Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. Cell *82*, 193-207.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

Zhang, Q.C., Deng, L., Fisher, M., Guan, J., Honig, B., and Petrey, D. (2011). PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res *39*, W283-287.

Zykovich, A., Korf, I., and Segal, D.J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing. Nucleic Acids Research *37*.